

Local Consistency of Markov Chain Monte Carlo Methods

Kengo KAMATANI^{*†}

Abstract

In this paper, we introduce the notion of efficiency (consistency) and examine some asymptotic properties of Markov chain Monte Carlo methods. We apply these results to the Gibbs sampler for independent and identically distributed observations. More precisely, we show that if both the sample size and the running time of the Gibbs sampler tend to infinity, and if the initial guess is not far from the true parameter, the empirical distribution of Gibbs sampler tends to a posterior distribution. This is a local property of the Gibbs sampler, which may be, in some cases, more essential than the global properties to describe its behavior. The advantages of using the local properties are the generality of the underlying model and the existence of simple equivalent Gibbs sampler. Those yield a simple regularity condition and suggest the reason for non-regular behaviors, which provides useful insight into the problem of how to construct efficient algorithms.

1 Introduction

This paper investigates conditions under which a Markov chain Monte Carlo (MCMC) method has a good stability property. There have a vast literature related to the sufficient conditions for ergodicity: see reviews [17] and [14] and textbooks such as [13] and [11]. The Markov probability transition of MCMC is Harris recurrent under fairly general assumptions. Moreover, it is sometimes geometrically ergodic. In practice, Foster-Lyapunov type drift conditions are commonly used to establish geometric ergodicity. This drift condition works well in studying MCMC stability, but there are some limitations.

- Technical difficulty in constructing a drift condition. See [3] for detail.
- The condition describes global properties of MCMC such as global convergence rate and global mixing rate, but not a local property. For some

^{*}Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan, kengok@ms.u-tokyo.ac.jp.

[†]Supported in part by Grant-in-Aid for JSPS Fellows (19-3140) and Grant-in-Aid for Young Scientists (B) 22740055.

MCMC methods, local properties seems important for describing MCMC efficiency/inefficiency (see ex. Examples 6 and 7 of [14]).

We take another approach to study stability of MCMC methods. It is well recognized that there are two kinds of randomness for Monte Carlo methods, such as Gibbs sampler. One is observation randomness and the other is simulation randomness. Usually we only consider the latter randomness for the analysis of Monte Carlo methods under fixed observation. However, it is natural to consider both randomness for the analysis, and in fact, the analysis becomes easier if we consider observation randomness. In particular, we can apply beautiful results of asymptotic statistics theory when we consider a large sample situation.

We obtain the following results.

1. Consistency and local consistency of Monte Carlo procedure are studied.
2. A reasonable set of sufficient conditions for consistency for the Gibbs sampler is addressed for independent identically distributed observations. We only assume (a) identifiability of parameter, (b) existence of uniformly consistent test, (c) regularity of prior distribution, and (d) quadratic mean differentiability of the full model.

The paper is divided into two parts. The first part is a study of Monte Carlo procedure, such as importance sampling and Gibbs sampler in general. We will describe Gibbs sampler as a sequence of random Monte Carlo procedure. We prepare in Section 2, a study of non-random Monte Carlo procedure and in Section 3 a study of sequence of non-random Monte Carlo procedure for the study of a sequence of random Monte Carlo procedure. Consistency and local consistency are introduced in Section 3. In Section 4, we consider Monte Carlo procedure in general including Gibbs sampler.

In the second part we consider more specific situation, more precisely, a large sample setting. In Section 5 we prepare some technical tools for the analysis of Gibbs sampler. In Section 6 we analyze local consistency of a sequence of standard Gibbs sampler in large sample setting.

For a treatment of a large sample setting (with a different motivation), a recent paper [2] studied the Metropolis algorithm for increased parameter dimension d . They obtained the rate of the running time of the Metropolis algorithm for burn-in and after burn-in. To deal with the complex algorithm and to obtain strong results, they assumed strong conditions (C.1, C.2 and (3.5)). Another paper, [12] and [16] obtained stability properties of the stochastic EM algorithm. Essentially they studied finite dimensional convergence of $\theta(0), \dots, \theta(k)$. However, without tightness arguments, the finite dimensional properties are insufficient to describe MCMC behaviors. On the other hand, we show the convergence of the law of the process $(\theta(i); i \in \mathbf{N}_0)$ with a minimal set of conditions.

It is not our intension to conclude that the Gibbs sampler is always efficient. The conclusion of Theorem 6.4 is that under a set of fairly general assumptions,

the empirical distribution constructed by the Gibbs sampler converges to the posterior distribution in a short running time. On the other hand, it illustrates the reason for the non-ideal behavior of the Gibbs sampler. For example, (a) we fail to take a good initial guess, (b) it fails to have a strong identifiability condition (c) the Fisher information matrix $I(\theta)$ is almost 0 or that for the hidden information $J(\theta)$ is almost ∞ , or (d) the sample size is too small related to its parameter dimension. For example, natural Gibbs sampler on probit regression model corresponds to the case (c) and it fails to hold good convergence property, which is studied in [7]. These studies of regular/non-regular properties of MCMC are important step constructing new efficient Monte Carlo algorithms including adaptive MCMC methods.

1.1 Notation

Let $\mathbf{N} = \{1, 2, \dots\}$ and $\mathbf{N}_0 = \{0, 1, 2, \dots\}$. We write the integer part of $x \in \mathbf{R}$ by $[x]$.

1.1.1 Probability measure, Transition kernel

For measurable space (E, \mathcal{E}) , the space of probability measures on (E, \mathcal{E}) is denoted by $\mathcal{P}(E)$.

For two measurable spaces (E, \mathcal{E}) and (F, \mathcal{F}) , a probability transition kernel K from E to F is a map $K : E \times \mathcal{F} \rightarrow [0, 1]$ such that

1. $K(x, \cdot)$ is a probability measure on (F, \mathcal{F}) for $x \in E$.
2. $K(\cdot, A)$ is \mathcal{E} -measurable for any $A \in \mathcal{F}$.

We may write $K(dy|x)$ instead of $K(x, dy)$. If $K(x, \cdot)$ is σ -finite measure instead of probability measure, we call K a transition kernel.

1.1.2 Normal distribution

Write $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ for a probability distribution function of $N(0, 1)$ and write $\Phi(x) = \int_{-\infty}^x \phi(y)dy$. For $\mu \in \mathbf{R}^p$ and $p \times p$ -positive definite matrix Σ , a function $\phi(x; \mu, \Sigma) = \exp(-x^T \Sigma^{-1} x/2)/(2\pi \det(\Sigma))^{1/2}$ is a probability distribution function of $N(\mu, \Sigma) = N_p(\mu, \Sigma)$ where $\det(\Sigma)$ is a determinant of Σ and x^T is a transpose of a vector $x \in \mathbf{R}^p$.

1.1.3 Centering

For a probability measure μ on \mathbf{R} , a central value is a point $\bar{x} \in \mathbf{R}$ satisfying

$$\int_{\mathbf{R}^p} \arctan(x - \bar{x}) \mu(dx) = 0.$$

Element of \mathbf{R}^p is denoted by $x = (x^1, \dots, x^p)^T$. For a probability measure μ on \mathbf{R}^p , let $\mu^i(A)$ be $\int_{x \in \mathbf{R}^p} 1_A(x^i) \mu(dx)$ for $A \in \mathcal{B}(\mathbf{R})$. For μ , we call $\bar{x} =$

$(\bar{x}^1, \bar{x}^2, \dots, \bar{x}^p)^T \in \mathbf{R}^p$ central value if each \bar{x}^i is a central value of μ^i . There is no practical reason for the use of the central value for Markov chain Monte Carlo procedure as is used in this paper. We use it because of its existence and continuity. That is, (a) for the posterior distribution $P_n(d\theta|x_n)$, its mean does not always exist but the central value does and moreover, it is unique and (b) if $\mu_n \rightarrow \mu$, then the central value of μ_n tends to that of μ . See [5].

2 Non-random Monte Carlo Procedure

Let (Θ, d) be a complete separable metric space equipped with Borel σ -algebra Ξ . Let (S, \mathcal{S}) be measurable space. Usually, $(S, \mathcal{S}) = (\Theta, \Xi)$ but it is not always the case. Write $S^{\mathbf{N}_0}$ for a countable product of S . We use a notation $s_\infty = (s(0), s(1), \dots) \in S^{\mathbf{N}_0}$. We write a subsequence $s_m = (s(0), s(1), \dots, s(m-1)) \in S^m$ of s_∞ .

2.1 Definition of non-random Monte Carlo Procedure

Now we are going to define a non-random Monte Carlo procedure. It may sound strange since Monte Carlo procedure has always randomness. The term “non-random” means that the Monte Carlo procedure does not depend on observation which will be denoted by x . We consider a Gibbs sampler as an example.

	observation randomness	simulation randomness
Non-random MC	X	O
(Random) MC	O	O

Gibbs sampler is a method which generate Markov chain $\theta_\infty = (\theta(0), \theta(1), \dots)$ depending on an observation x . Thus there are two kinds of randomness induced by x and θ_∞ . The former randomness is an observation randomness and the latter is a simulation randomness. Any Monte Carlo procedure uses a simulation randomness but it may not use observation randomness. We call it non-random if it does not use observation randomness.

Non-random Monte Carlo procedure is constructed by a probability measure M on $S^{\mathbf{N}_0}$ and a sequence $e = (e_m; m = 1, 2, \dots)$ where e_m is a probability transition kernel from S^m to Θ . Now we consider a simple example for the case $\Theta = S$. If we approximate integral of measurable function f of Θ with respect to a probability measure Π on Θ , that is,

$$\Pi(f) = \int_{x \in \Theta} f(x) \Pi(dx)$$

we generate independent sequence $\theta_m = (\theta(0), \dots, \theta(m-1))$ from Π and set

$$\frac{1}{m} \sum_{i=0}^{m-1} f(\theta(i)). \quad (2.1)$$

This procedure is generated by $M = \Pi^{\otimes \mathbb{N}_0}$ (see Example 2.2) with

$$e_m(\theta_m, \cdot) = \frac{1}{m} \sum_{i=0}^{m-1} \delta_{\theta(i)}$$

where δ_θ is a Dirac measure with its mass on $\theta \in \Theta$. Then (2.1) is $\int_\Theta f(\theta) e_m(\theta_m, d\theta)$. Thus we define Monte Carlo procedure as follows.

Definition 2.1 (Non-random Monte Carlo procedure). *Let M be a probability measure on $S^{\mathbb{N}_0}$ and let e_m be a probability transition kernel from S^m to Θ for $m = 1, 2, \dots$ and set $e = (e_m; m = 1, 2, \dots)$. We call $\mathcal{M} = (M, e)$ a non-random Monte Carlo procedure on (S, Θ) .*

If $S = \Theta$, we call \mathcal{M} , a non-random Monte Carlo procedure on Θ . Before introducing examples, we prepare some remarks. For simplicity, we write $e_m(s_\infty)$ or $e_m(s_m)$ instead of $e_m(s_m, \cdot)$. When $S = \Theta$, if $e = (e_m; m = 1, 2, \dots)$ is defined by (2.1) we call e a sequence of empirical distribution. If (S, \mathcal{S}) is a product of $(\Theta, \Xi) \times (Y, \mathcal{Y})$ for some measurable space (Y, \mathcal{Y}) and if $e_m(s_m)$ ($s_m = ((\theta(0), y(0)), \dots, (\theta(m-1), y(m-1)))$) is the same as the right hand side of (2.1), then we call e a sequence of empirical distribution on Θ .

Many Monte Carlo methods can be represented as a Monte Carlo procedure defined above. An important exception is some sequential Monte Carlo methods. It is felt to require other framework for the analysis of sequential Monte Carlo methods.

Example 2.2 (Non-random crude Monte Carlo procedure). *Take $S = \Theta$. Let Π be a probability measure on Θ and e be a sequence of empirical distribution. Let $M = \Pi^{\otimes \mathbb{N}_0}$, that is,*

$$M(d\theta_\infty) = \prod_{i=0}^{\infty} \Pi(d\theta(i)).$$

Then a non-random Monte Carlo procedure $\mathcal{M} = (M, e)$ is called a non-random crude Monte Carlo procedure on Θ .

Example 2.3 (Non-random importance sampling procedure). *Take $S = \Theta$. Let Π, Q be a probability measure on Θ and let Π be absolutely continuous with respect to Q , that is, $\Pi(A) = 0$ if $Q(A) = 0$ for $A \in \Xi$. Let $e = (e_m; m = 1, 2, \dots)$ be*

$$e_m(\theta_m) = \frac{1}{m} \sum_{i=0}^{m-1} \frac{d\Pi}{dQ}(\theta(i)) \delta_{\theta(i)}$$

and $M = Q^{\mathbb{N}_0}$. Then a non-random Monte Carlo procedure $\mathcal{M} = (M, e)$ is called a non-random importance sampling procedure on Θ .

Non-random accept-reject method has a similar form. All of these Monte Carlo procedure bases on countable products of probability measures $M = \Pi^{\otimes \mathbb{N}_0}$. The properties of these Monte Carlo procedure will be discussed later.

2.2 Non-random Markov chain Monte Carlo procedure

We are going to define non-random Markov chain Monte Carlo procedure as a class of non-random Monte Carlo procedure. Let μ be a probability measure on S and K be a probability transition kernel on S . Then

$$M(ds_\infty) = \mu(ds(0)) \prod_{i=0}^{\infty} K(s(i), ds(i+1))$$

is said to be Markov measure on $S^{\mathbb{N}_0}$ generated by (μ, K) .

Definition 2.4 (Non-random Markov chain Monte Carlo procedure). *Let M be a Markov measure on $S^{\mathbb{N}_0}$ and let e_m be a probability transition kernel from S^m to Θ for $m = 1, 2, \dots$ and set $e = (e_m; m = 1, 2, \dots)$. We call $\mathcal{M} = (M, e)$ a non-random Markov chain Monte Carlo procedure on (S, Θ) .*

First we state some possibilities for e .

Example 2.5 (Burn-inn, thinning). *Take $S = \Theta$. Let $\mathcal{M} = (M, e)$ be non-random Markov chain Monte Carlo procedure. If $e = (e_m; m = 1, 2, \dots)$ is defined by*

$$e_m(s_m) = \frac{1}{m - [m/2]} \sum_{i=[m/2]}^{m-1} \delta_{\theta(i)}$$

e is called a sequence of empirical distribution with burn-inn, and if

$$e_m(\theta_m) = \frac{1}{[m/2]} \sum_{i=0}^{[m/2]-1} \delta_{\theta(2i)},$$

e is called a sequence of empirical distribution with thinning.

Now we are going to define a Gibbs sampler as an example of non-random Monte Carlo procedure.

Example 2.6 (Non-random Gibbs sampler). *Let (Y, \mathcal{Y}) be a measurable space and set $(S, \mathcal{S}) = (Y, \mathcal{Y}) \otimes (\Theta, \Xi)$. Let $P(dy|\theta)$ and $P(d\theta|y)$ be a probability transiting kernel from Θ to Y and Y to Θ with respectively. Let $\overline{\mathcal{M}} = (\overline{M}, \overline{e})$ be a Markov chain Monte Carlo procedure on (S, Θ) having \overline{K} as a probability transition kernel for \overline{M} defined by*

$$\overline{K}((y, \theta), d(y^*, \theta^*)) = P(dy^*|\theta)P(d\theta^*|y^*).$$

Then $\overline{\mathcal{M}}$ is called a non-random Gibbs sampler on (S, Θ) . Note that $\overline{K}((y, \theta), \cdot)$ does not depend on y .

For the analysis of the Gibbs sampler, usually, it is sufficient to study

$$K(\theta, d\theta^*) = \int_{y \in Y} P(dy|\theta)P(d\theta^*|y). \quad (2.2)$$

See Definition 2.12 for detail. An important exception, which requires the analysis of \bar{K} instead of K is Rao-Blackwellization strategy. Rao-Blackwellization is an effective strategy for Markov chain Monte Carlo procedure. We consider this strategy as a one example of e .

Example 2.7 (Rao-Blackwellization). *Let $\bar{\mathcal{M}}$ be a non-random Gibbs sampler on (S, Θ) defined as above. Take \bar{e}_m as*

$$\bar{e}_m(s_m, A) = \frac{1}{m} \sum_{i=0}^{m-1} \int_{\theta \in A} P(d\theta | y(i))$$

where $s_m = (s(0), \dots, s(m-1))$, $s(i) = (y(i), \theta(i))$ for $A \in \Xi$. Then for the choice of $\bar{e} = (\bar{e}_m; m = 1, \dots)$, \bar{e} is called an empirical distribution with Rao-Blackwellization.

Next we define Metropolis-Hastings algorithm in the following example. Transition kernel \bar{K} defined in the following example may be different from usual one denoted by K in (2.4). We will explain the relation of two transition kernels after the following example.

Example 2.8 (Non-random Metropolis-Hastings procedure). *Let $S = \Theta \times [0, 1] \times \Theta$. Let Π be a probability measure on Θ and Q be a probability transition kernel on S . Let r be a Ξ^2 -measurable function such that*

$$r(x, y)\Pi(dx)Q(x, dy) = \Pi(dy)Q(y, dx).$$

Let $\alpha(x, y) = \min\{1, r(x, y)\}$ be a measurable function called acceptance ratio. We define a probability transition kernel $\bar{K}((x, u, y), d(x^*, u^*, y^*))$ from S to itself by

$$Q(y, dx^*)1_{[0,1]}(u)du(1(u \leq \alpha(y, x^*))\delta_{x^*}(dy^*) + 1(u > \alpha(y, x^*))\delta_y(dy^*)) \quad (2.3)$$

When $\bar{\mathcal{M}} = (\bar{\mathcal{M}}, \bar{e})$ is a non-random Markov chain Monte Carlo procedure and $\bar{\mathcal{M}}$ has a probability transition kernel \bar{K} , we call $\bar{\mathcal{M}}$ non-random Metropolis-Hastings procedure on (S, Θ) generated by (Π, Q) . Note that $\bar{K}((x, u, y), \cdot)$ does not depend on x, u .

The above representation (2.3) of the transition kernel shows all realization (a) propose x , (b) $u \sim U[0, 1]$, (c) y result of accept-reject procedure. When we are only interested in y , we can use simpler notation which is a usual one. For $A(x) = \int_{y \in S} \alpha(x, y)Q(x, dy)$, the following transition kernel is simpler:

$$K(x, dy) = \alpha(x, y)Q(x, dy) + (1 - A(x))\delta_x(dy). \quad (2.4)$$

As the Gibbs sampler, usually it is sufficient to consider K (see Definition 2.12), but not always the case. Some algorithm such as [1] uses the information of proposed variable x .

We define ergodicity and stationarity for a non-random Monte Carlo procedure. We call $\mathcal{M} = (M, e)$ ergodic or stationary if M is ergodic or stationary. Recall some terminology related to ergodicity and stationarity (see monographs such as [4] or [15]). Let $T(s_\infty) = (s(1), s(2) \dots)$ ($s_\infty = (s(0), s(1), \dots)$).

- Probability measure M is said to be (strictly) stationary if $M(A) = M(T^{-1}A)$. When M is stationary, a probability measure Π defined by $\Pi(A) := M(\{s_\infty; s(0) \in A\})$ ($A \in \mathcal{S}$) is called invariant probability measure.
- A set $A \in \mathcal{S}^{\mathbb{N}_0}$ is called invariant if $T^{-1}A = A$. Let \mathcal{A} be a σ -algebra generated by the invariant sets.
- M is called ergodic if $M(A) = 0$ or 1 for any $A \in \mathcal{A}$.

If M is stationary and ergodic, we have the ergodic theorem (see Theorem 10.2.1 of [4]). If M is a Markov measure generated by irreducible and positive Harris recurrent probability transition kernel, M is ergodic.

Definition 2.9. Let $\mathcal{M} = (M, e)$ be a non-random Monte Carlo procedure. When M is ergodic or stationary, we call \mathcal{M} ergodic or stationary with respectively.

Stationarity and ergodicity play an important role for convergence of Monte Carlo procedure.

2.3 Consistency of non-random Monte Carlo procedure

Let BL_1 be a class of Ξ -measurable \mathbf{R} -valued functions f satisfying

$$|f(s) - f(t)| \leq d(s, t) \quad (s, t \in \Theta).$$

When μ, ν are probability measures on Θ , let $w(\mu, \nu)$ denote the bounded Lipschitz metric, that is,

$$w(\mu, \nu) = \sup_{\psi \in \text{BL}_1} \left| \int_{x \in \Theta} \psi(x) \mu(dx) - \int_{x \in \Theta} \psi(x) \nu(dx) \right|$$

This metric is equivalent to the Prohorov metric, that is $w(\mu_n, \nu) \rightarrow 0$ is equivalent to the weak convergence. We may write $\text{BL}_1(\Theta)$ and w_Θ instead of BL_1 and w to indicate the underlying space.

For Monte Carlo procedure $\mathcal{M} = (M, e)$ and a probability measure Π on Θ , we define a risk function

$$R_m(\mathcal{M}, \Pi) = \int_{s_\infty \in S^{\mathbb{N}_0}} w(e_m(s_\infty), \Pi) M(ds_\infty).$$

Definition 2.10 (Consistency). Let $\mathcal{M} = (M, e)$ be a non-random Monte Carlo procedure on (S, Θ) and Π be a probability measure on Θ . Then \mathcal{M} is said to be consistent to Π if $\lim_{m \rightarrow \infty} R_m(\mathcal{M}, \Pi) = 0$.

Proposition 2.11. Non-random crude Monte Carlo procedure and non-random importance sampling procedure is consistent to Π . Moreover, non-random Monte Carlo procedure is consistent to Π if M is stationary and ergodic with invariant probability measure Π and e is a sequence of empirical distribution.

Proof. In each case, the weak law of large numbers hold, that is, for any Π -integrable function f

$$\int_{\theta \in \Theta} f(\theta) e_m(s_m)(d\theta) - \int_{\theta \in \Theta} f(\theta) \Pi(d\theta) \rightarrow 0$$

in M -probability. For $\epsilon > 0$, by separability and complicity of Θ , there is a relatively compact open set K of Θ satisfying $\Pi(K^c) \leq \epsilon/4$. On a relatively compact set K , we can choose a finite sequence $\psi_1, \dots, \psi_k \in \text{BL}_1$ such that for any $\psi \in \text{BL}_1$, there exists $i \in \{1, \dots, k\}$ such that

$$\sup_{s \in K} |\psi(s) - \psi_i(s)| \leq \epsilon/2.$$

Let $\xi_n : \Theta \rightarrow [0, 1]$ be a bounded Lipschitz function such that $\xi_n \downarrow 1_{K^c}$. By taking $\psi = \psi \xi_n + (\psi - \psi_i)(1 - \xi_n) + \psi_i(1 - \xi_n)$ and write $\psi_i(1 - \xi_n) =: \psi_{i,n}$, $w(e_m(s_m), \Pi)$ is bounded above by

$$e_m(s_m, \xi_n) + \Pi(\xi_n) + \frac{\epsilon}{2} + \sum_{i=1}^k \left| \int \psi_{i,n}(\theta) e_m(s_m)(d\theta) - \int \psi_{i,n}(\theta) \Pi(d\theta) \right|,$$

where $e_m(s_m, \xi_n)$ and $\Pi(\xi_n)$ are integrals of ξ_n with respect to $e_m(s_m, \cdot)$ and Π . The M -integral of the last term tends to 0, and the first term tends to $\Pi(\xi_n)$ as $m \rightarrow \infty$ since $\int_{S_{\mathbf{N}_0}} M(ds_\infty) e_m(s_m, \cdot)$ converges weakly to Π . Then taking $n \rightarrow \infty$, $\limsup_{m \rightarrow \infty} R_m(\mathcal{M}, \Pi)$ is bounded above by

$$\Pi(K^c) + \Pi(K^c) + \frac{\epsilon}{2} + 0 \leq \epsilon.$$

Hence $\limsup_{m \rightarrow \infty} R_m(\mathcal{M}, \Pi) = 0$ as required. \square

We define equivalence of Monte Carlo procedures already mentioned in Examples 2.6 and 2.8.

Definition 2.12 (Equivalence). *Let (Θ, d) be metric space with Borel σ -algebra Ξ , and let (S^i, \mathcal{S}^i) be measurable spaces for $i = 1, 2$. Let $\mathcal{M}^i = (M^i, e^i)$ be Monte Carlo procedure on (S^i, Θ) for $i = 1, 2$. Then \mathcal{M}^1 and \mathcal{M}^2 are called equivalent if*

$$R_m(\mathcal{M}^1, \Pi) = R_m(\mathcal{M}^2, \Pi)$$

for any $m \in \mathbf{N}$ and probability measure Π on Θ .

Example 2.13. Assume the same condition as Example 2.6. Let K be as in (2.2) and $\mu(d\theta) = \int_{y \in Y} \bar{\mu}(dy d\theta)$ where $\bar{\mu}$ is the initial probability measure of \bar{M} and set M as a Markov measure defined by μ and K . If \bar{e}_m does not depend on Y^m for $m = 1, 2, \dots$, then $\mathcal{M} = (M, e)$ is equivalent to $\bar{\mathcal{M}}$ where $e = (e_i; i = 1, 2, \dots)$ and $e_m(\theta_m) = \bar{e}_m(s_m)$.

3 Sequence of non-random Monte Carlo procedure

In this section, we consider a sequence of non-random Monte Carlo procedure. Therefore we will consider a sequence of measurable spaces (S_n, \mathcal{S}_n) and a sequence of complete and separable metric spaces (Θ_n, d^n) . Write w_m for the bounded Lipschitz metric for the space of probability measures on Θ_n , corresponding to the metric d^n .

Throughout in this section, $\mathcal{M}_n = (M_n, e_n)$ is a non-random Monte Carlo procedure on (S_n, Θ_n) where M_n is a probability measure on $S_n^{\mathbf{N}_0}$ and $e_n = (e_{n,m}; m = 1, 2, \dots)$ is a sequence of probability transition kernels $e_{n,m}$ from S_n^m to Θ_n for $m = 1, 2, \dots$. We write $s_{n,m} = (s_n(0), s_n(1), \dots, s_n(m-1)) \in S_n^m$ and $s_{n,\infty} = (s_n(0), s_n(1), \dots) \in S_n^{\mathbf{N}_0}$. As in the previous section, we may write $e_{n,m}(s_{n,\infty})$ instead of $e_{n,m}(s_{n,m}, \cdot)$.

3.1 Consistency of sequence of non-random Monte Carlo procedure

Let Π_n be a probability measure on Θ_n for each $n = 1, 2, \dots$. We define a risk function for each $\mathcal{M}_n = (M_n, e_n)$ for $e_n = (e_{n,m}; m = 1, 2, \dots)$ by

$$R_m(\mathcal{M}_n, \Pi_n) = \int_{s_{n,\infty} \in S_n^{\mathbf{N}_0}} w_n(e_{n,m}(s_{n,\infty}), \Pi_n) M_n(ds_{n,\infty}).$$

Definition 3.1 (Consistency for sequence). *For $n = 1, 2, \dots$, $\mathcal{M}_n = (M_n, e_n)$ is a non-random Monte Carlo procedure on (S_n, Θ_n) and Π_n is a probability measure on Θ_n . Then $(\mathcal{M}_n; n = 1, 2, \dots)$ is said to be consistent to $(\Pi_n; n = 1, 2, \dots)$ if $\lim_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) = 0$ for any $m_n \rightarrow \infty$.*

We show some non-consistent examples (see also [7] for other type of non-consistency, degeneracy). First example is an importance sampling in high dimension.

Example 3.2. *Let $\Theta_n = S_n = \mathbf{R}^n$ and let I_n be the n -dimensional identity matrix. Let $N_n(\mu_n, \Sigma_n)$ be a normal distribution with mean $\mu_n \in \mathbf{R}^n$ and positive definite matrix Σ_n . Consider two probability measures $\Pi_n = N_n(0, I_n)$ and $Q_n = N_n((1, 0, \dots, 0)^T, I_n)$ and let $\mathcal{M}_n = (M_n, e_n)$ be non-random crude Monte Carlo procedure. We show that $(\mathcal{M}_n; n = 1, 2, \dots)$ is not consistent to $(\Pi_n; n = 1, 2, \dots)$. Denote $\theta = (\theta^1, \dots, \theta^n)^T \in \Theta_n$. Each projection of Q_n to i -th coordinate θ^i is $N(1, 1)$ for $i = 1$ and $N(0, 1)$ for $i = 2, \dots, n$. Let*

$$N_{n,m}^i = \{(\theta_n(0), \theta_n(1), \dots); \theta_n^i(j) < 0 \ (j = 0, \dots, m-1)\}$$

where $\theta_n(i) = (\theta_n^1(i), \dots, \theta_n^n(i))^T \in \Theta_n$. The event $N_{n,m}^i$ has probability 2^{-m} under M_n for $i = 2, \dots, n$. Therefore, by independence of the events $(N_{n,m}^i; i = 2, \dots, n)$, $N_{n,m} := \cup_{i=2}^n N_{n,m}^i$ has probability $1 - (1 - 2^{-m})^{n-1}$. For $i = 2, \dots, n$, take $\psi \in \text{BL}_1$ to be

$$\psi(x) = \max\{0, \min\{1, x^i\}\} \ (x = (x^1, \dots, x^n)^T \in \Theta_n).$$

Then if $(\theta_n(0), \theta_n(1), \dots) \in N_{n,m}^i$, we have

$$|\int_{x \in \Theta_n} \psi(x) e_{n,m}(\theta_{n,m})(dx) - \int_{x \in \Theta_n} \psi(x) \Pi_n(dx)| = \int_0^\infty \min\{1, x\} \phi(x) dx =: c > 0.$$

Therefore $w_n(e_{n,m}(\theta_{n,m}), \Pi_n) \geq c$ on $N_{n,m}$ and hence

$$R_m(\mathcal{M}_n, \Pi_n) \geq (1 - (1 - 2^{-m})^{n-1})c.$$

We can choose $m_n \rightarrow \infty$ to be $\liminf_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) > 0$. Hence it is not consistent.

(Geometric, or uniform) Ergodicity may not provide enough information whether a given Markov chain Monte Carlo method works well or not. In that approach, we have to analyze good estimate of the convergence rate of the total variation distance of the marginal distribution or asymptotic variance of the empirical mean. The analysis of consistency may provide another viewpoint. Sometimes it provides a good information for the behavior of Markov chain Monte Carlo methods.

Example 3.3. Let $\Theta_n = [-n, n]$ and $S_n = \Theta_n \times [0, 1] \times \Theta_n$. For $n = 1, 2, \dots$ let Π_n be a restriction of $N(0, 1)$ to the interval $[-n, n]$. Let $Q_n(x, dy) \equiv Q(dy) = N(0, n^{-1})$. Consider non-random Metropolis-Hastings procedure $\overline{\mathcal{M}}_n = (\overline{M}_n, \overline{e}_n)$ on Θ_n generated by (Π_n, Q_n) where \overline{e}_n is a sequence of empirical distribution on Θ and M_n is a Markov measure with initial distribution $\delta_0(dx)1_{[0,1]}(du)Q(dy)$.

Intuitively, this non-random Metropolis-Hastings procedure works poorly, and it is true. It is easy to see by checking consistency and degeneracy

Consider equivalent Markov chain Monte Carlo procedure $\mathcal{M}_n = (M_n, e)$ as in the comment after Example 2.8. For any fixed $m \in \mathbb{N}$,

$$M_n(\{\theta_\infty; \max_{i=0, \dots, m-1} |\theta(i)| > 1\}) \leq Q^{\otimes N_0}(\{\theta_\infty; \max_{i=0, \dots, m-1} |\theta(i)| > 1\}) \rightarrow 0$$

for $\theta_\infty = (\theta(0), \theta(1), \dots)$. Take $\psi \in \text{BL}_1$ to be

$$\psi(x) = \max\{0, \min\{1, |x| - 1\}\} \quad (x \in \Theta_n).$$

Then if $\max_{i=0, \dots, m-1} |\theta(i)| \leq 1$,

$$|\int_{x \in S_n} \psi(x) e_m(\theta_m)(dx) - \int_{x \in S_n} \psi(x) \Pi_n(dx)| = \int_{|x| > 1} \frac{\min\{1, |x| - 1\}}{1 - 2\Phi(n)} \psi(x) dx > c > 0$$

Then

$$R_m(\mathcal{M}_n, \Pi_n) \geq M_n(\{\theta_\infty; \max_{i=0, \dots, m-1} |\theta(i)| \leq 1\})c \rightarrow c.$$

Hence we can choose $m_n \rightarrow \infty$ to be $\liminf_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) > 0$. Therefore $(\overline{\mathcal{M}}_n; n = 1, 2, \dots)$ is not consistent to $(\Pi_n; n = 1, 2, \dots)$.

3.2 Sufficient conditions for consistency of non-random Markov chain Monte Carlo procedure

Let (Θ, d) be a complete separable metric space equipped with Borel σ -algebra Ξ . In this subsection, we assume

$$(\Theta_n, d^n) \equiv (\Theta, d), \quad (S_n, \mathcal{S}_n) \equiv (\Theta, \Xi), \quad d(s, t) \leq 1 \quad (s, t \in \Theta). \quad (3.1)$$

Note that the assumption $d(s, t) \leq 1$ is just for simplicity and all results in this paper are valid without this assumption.

Write $\Theta^{\mathbf{N}_0}$ for a countable product of Θ . We use a notation $\theta_\infty = (\theta(0), \theta(1), \dots) \in \Theta^{\mathbf{N}_0}$. We introduce a metric d_∞ by

$$d_\infty(\theta_\infty^1, \theta_\infty^2) = \sum_{i=0}^{\infty} 2^{-i-1} d(\theta^1(i), \theta^2(i)) \quad (3.2)$$

where $\theta_\infty^i = (\theta^i(0), \theta^i(1), \dots) \in \Theta^{\mathbf{N}_0}$ ($i = 1, 2$). We write a subsequence $\theta_m = (\theta(0), \theta(1), \dots, \theta(m-1)) \in \Theta^m$ of θ_∞ introducing a metric d_m such that $d_m(\theta_m^1, \theta_m^2)$ is the same as the right hand side of (3.2) replacing “ ∞ ” by “ $m-1$ ” where θ_m^i is subsequence of θ_∞^i for first m elements. Let w_∞ and w_m be a bounded Lipschitz metric for $\mathcal{P}(\Theta^{\mathbf{N}_0})$ and $\mathcal{P}(\Theta^m)$ defined by d_∞ and d_m with respectively. The next two propositions are fundamental results for the consistency of Monte Carlo procedure.

Proposition 3.4. *Let $\mathcal{M}_n = (M_n, e)$ be a non-random stationary Monte Carlo procedure with invariant distribution Π_n for $n = 1, 2, \dots, \infty$. Moreover, \mathcal{M}_∞ is ergodic and e is a sequence of empirical distribution. If $w_\infty(M_n, M_\infty) \rightarrow 0$, then $(M_n; n = 1, 2, \dots)$ is consistent to $(\Pi_n; n = 1, 2, \dots)$.*

Proof. First we show that for stationary Monte Carlo procedure $\mathcal{M} = (M, e)$, and for $k \leq m$,

$$R_m(\mathcal{M}, \Pi) \leq R_k(\mathcal{M}, \Pi) + \frac{k}{m}. \quad (3.3)$$

Write $\Pi(\psi)$ for $\int \psi(x) \Pi(dx)$. By definition,

$$w(e_m(\theta_m), \Pi) = \sup_{\psi \in \text{BL}_1} \left| \frac{1}{m} \sum_{i=0}^{m-1} \psi(\theta(i)) - \Pi(\psi) \right|.$$

We divide sequence θ_m into subsequence of length k , that is, divide θ_m into $\theta_k^j = (\theta(jk), \theta(jk+1), \dots, \theta((j+1)k-1))$ ($j = 0, \dots, [m/k]-1$) and $\theta(k[m/k]), \dots, \theta(m-1)$. Then

$$\frac{1}{m} \sum_{i=0}^{m-1} \psi(\theta(i)) = \frac{k}{m} \sum_{j=0}^{[m/k]-1} \frac{1}{k} \sum_{i=0}^{k-1} \psi(\theta(jk+i)) + \frac{1}{m} \sum_{i=k[m/k]}^{m-1} \psi(\theta(i)).$$

This relation yields

$$w(e_m(\theta_m), \Pi) \leq \frac{k}{m} \sum_{j=0}^{[m/k]-1} w(e_k(\theta_k^j), \Pi) + \frac{1}{m} \sum_{i=k[m/k]}^{m-1} |\psi(\theta(i)) - \Pi(\psi)|.$$

For the first term of the right hand side, by stationarity, each $w(e_k(\theta_k^j), \Pi)$ for $j = 0, \dots, [m/k] - 1$ has the same law under M . For the second term, we have a relation

$$|\psi(\theta) - \Pi(\psi)| \leq \int_{\theta^* \in \Theta} |\psi(\theta) - \psi(\theta^*)| \Pi(d\theta^*) \leq \int d(\theta, \theta^*) \Pi(d\theta^*) \leq 1.$$

Using these relations, we have

$$R_m(\mathcal{M}, \Pi) \leq \frac{k}{m} \left[\frac{m}{k} \right] R_k(\mathcal{M}, \Pi) + \frac{m - k[m/k]}{m}.$$

Since $x - 1 < [x] \leq x$, (3.3) follows. Applying this result to \mathcal{M}_n and Π_n , we have for any $m_n \rightarrow \infty$,

$$\limsup_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) \leq \limsup_{n \rightarrow \infty} R_k(\mathcal{M}_n, \Pi_n) \leq \limsup_{n \rightarrow \infty} R_k(\mathcal{M}_n, \Pi_\infty)$$

where the second inequality comes from $R_k(\mathcal{M}_n, \Pi_n) \leq R_k(\mathcal{M}_n, \Pi_\infty) + w(\Pi_n, \Pi_\infty)$ and $w(\Pi_n, \Pi_\infty) \rightarrow 0$ by $w_\infty(M_n, M_\infty) \rightarrow 0$.

Now we are going to show the continuity of $\theta_\infty \mapsto w(e_k(\theta_\infty), \Pi_\infty)$. If we have the property, $\limsup_{n \rightarrow \infty} R_k(\mathcal{M}_n, \Pi_\infty) = R_k(\mathcal{M}_\infty, \Pi_\infty)$ and it tends to 0 as $k \rightarrow \infty$ by Proposition 2.11. Therefore it is sufficient to show the continuity for the proof of $R_{m_n}(\mathcal{M}_n, \Pi_n) \rightarrow 0$. For $\theta_\infty^i = (\theta^i(0), \theta^i(1), \dots)$ ($i = 1, 2$), by triangular inequality,

$$|w(e_k(\theta_\infty^1), \Pi_\infty) - w(e_k(\theta_\infty^2), \Pi_\infty)| \leq |w(e_k(\theta_\infty^1), e_k(\theta_\infty^2))| = \sup_{\psi \in \text{BL}_1} \left| \frac{1}{m} \sum_{i=0}^{m-1} \psi(\theta^1(i)) - \psi(\theta^2(i)) \right|$$

which is bounded above by

$$\frac{1}{m} \sum_{i=0}^{m-1} d(\theta^1(i), \theta^2(i)) \leq 2^m d_\infty(\theta_\infty^1, \theta_\infty^2).$$

Hence $w(e_k(\theta_\infty), \Pi_\infty)$ is continuous and the claim follows. \square

Let μ be a probability measure on Θ and let K be probability transition kernel on Θ . Let $\mu \otimes K$ be a probability measure on Θ^2 defined by

$$\mu \otimes K(d\theta, d\theta^*) = \mu(d\theta) K(\theta, d\theta^*).$$

For any probability measures p, q on a metric space (E, d) with Borel σ -algebra \mathcal{E} , we define total variation distance by

$$\|p - q\| = \sup_{A \in \mathcal{E}} |p(A) - q(A)| \geq w_E(p, q) \quad (3.4)$$

where w_E is a bounded Lipschitz metric on the space of probability measures on E .

The following lemma due to Le Cam is very useful for our purpose. See Lemma 12.2.2 of [8] or Lemma 6.4.2 of [9].

Lemma 3.5. *Let (Θ, d) be a separable and complete metric space with Borel σ -algebra Ξ . Let μ_i be a probability measure on Θ and K_i be a probability transition kernel from Θ to itself for $i = 1, 2$. Then*

$$\int_{x \in \Theta} \|K_1(x, \cdot) - K_2(x, \cdot)\|(\mu_1 + \mu_2)(dx) \leq 4\|\mu_1 \otimes K_1 - \mu_2 \otimes K_2\|.$$

Using this lemma, the following is just an easy corollary.

Proposition 3.6. *Let $\mathcal{M}_n^i = (M_n^i, e_n^i)$ be non-random stationary Markov chain Monte Carlo procedure with invariant probability distributions Π_n^i for $i = 1, 2$ and $n = 1, 2, \dots$. Write K_n^i for the probability transition kernel of M_n^i . If $\|\Pi_n^1 \otimes K_n^1 - \Pi_n^2 \otimes K_n^2\| \rightarrow 0$, then $w_\infty(M_n^1, M_n^2) \rightarrow 0$.*

Proof. Write $M_{n,m}^i$ for the restriction of M_n^i to (Θ^m, Ξ^m) , that is,

$$M_{n,m}^i(d\theta_m) = \Pi_n^i(d\theta(0)) \prod_{i=0}^{m-2} K_n^i(\theta(i), d\theta(i+1)) \quad (\theta_m = (\theta(0), \dots, \theta(m-1))).$$

Recall that w_m is a bounded Lipschitz metric for the space of probability measures of Θ^m . Write $(\theta_m, 0) \in \Theta^{\mathbf{N}_0}$ for $(\theta(0), \dots, \theta(m-1), 0, 0, \dots)$ where 0 means just a fixed element of Θ . By definition

$$w_\infty(M_n^1, M_n^2) = \sup_{\psi \in \text{BL}_1} \left| \int \psi(\theta_\infty) M_n^1(d\theta_\infty) - \int \psi(\theta_\infty) M_n^2(d\theta_\infty) \right|$$

and by taking $\psi(\theta_\infty) = \psi(\theta_m, 0) + (\psi(\theta_\infty) - \psi(\theta_m, 0))$ and by $|\psi(\theta_\infty) - \psi(\theta_m, 0)| \leq d_\infty(\theta_\infty, (\theta_m, 0)) \leq 2^{-m}$ the above is bounded above by

$$w_m(M_{n,m}^1, M_{n,m}^2) + 2 * 2^{-m}.$$

Therefore, to show $w_\infty(M_n^1, M_n^2) \rightarrow 0$, it is sufficient to show $w_m(M_{n,m}^1, M_{n,m}^2) \rightarrow 0$ for any $m = 1, 2, \dots$. In fact, we can show $\|M_{n,m}^1 - M_{n,m}^2\| \rightarrow 0$ for any $m = 1, 2, \dots$ which is stronger than $w_m(M_{n,m}^1, M_{n,m}^2) \rightarrow 0$ by (3.4).

The convergence holds for $m = 1, 2$ by assumption. Now assume that the convergence is true for any $m = 1, 2, \dots, k$. For $m = k + 1$, observe that $(M_{n,k+1}^1 - M_{n,k+1}^2)(d\theta_{k+1})$ equals to

$$(M_{n,k}^1 - M_{n,k}^2)(d\theta_k) K_n^1(\theta(k), d\theta(k+1)) + M_{n,k}^2(d\theta_k)(K_n^1 - K_n^2)(\theta(k), d\theta(k+1)).$$

Then $\|M_{n,k+1}^1 - M_{n,k+1}^2\|$ is bounded above by

$$\|M_{n,k}^1 - M_{n,k}^2\| + \int M_{n,k}^2(d\theta_k) \|(K_n^1 - K_n^2)(\theta(k), \cdot)\|.$$

The former tends to 0 by the assumption of the induction. Since (Θ, d) is separable and complete, the second term equals to

$$\int \Pi_n^2(d\theta) \|(K_n^1 - K_n^2)(\theta, \cdot)\| \leq 4\|\Pi_n^1 \otimes K_n^1 - \Pi_n^2 \otimes K_n^2\| \rightarrow 0$$

by Lemma 3.5. □

3.3 Localization

In this section, we consider localization of non-random Monte Carlo procedure. The following example illustrates its motivation.

Example 3.7. Let $\Theta_n = S_n = \mathbf{R}$ and $\Pi_n = N(0, n^{-1})$. For $n = 1, 2, \dots$, let $\mathcal{M}_{1,n}$ and $\mathcal{M}_{2,n}$ be non-random crude Monte Carlo procedures corresponding to

$$Q_{1,n} = \delta_{\{0\}}, \quad Q_{2,n} = N(n^{-1}, n^{-1}),$$

with respectively with sequence of empirical distribution e . In the comment after Proposition 3.9, we will show that both $(\mathcal{M}_{1,n}; n = 1, 2, \dots)$ and $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ are consistent to $(\Pi_n; n = 1, 2, \dots)$. However the latter seems preferable than the former.

Make a projection $\varphi : \theta \mapsto n^{1/2}\theta$. Then the probability measures becomes $\Pi_n^* \equiv \Pi^* = N(0, 1)$ and

$$Q_{1,n}^* = \delta_{\{0\}}, \quad Q_{2,n}^* = N(n^{-1/2}, 1).$$

Let $\mathcal{M}_{1,n}^*$ and $\mathcal{M}_{2,n}^*$ be corresponding non-random crude Monte Carlo. Then $(\mathcal{M}_{1,n}^*; n = 1, 2, \dots)$ is not consistent to $(\Pi_n^*; n = 1, 2, \dots)$ since

$$R_m(\mathcal{M}_{1,n}^*, \Pi_n^*) = w(\delta_0, N(0, 1)) > 0 \quad (m, n = 1, 2, \dots).$$

On the other hand, if we write non-random crude Monte Carlo for Π^* with e by \mathcal{M}_0^* , then

$$R_{m_n}(\mathcal{M}_{2,n}^*, \Pi_n^*) = R_{m_n}(\mathcal{M}_{2,n}^*, N(n^{-1/2}, 1)) + o(1) = R_{m_n}(\mathcal{M}_0^*, \Pi^*) + o(1) = o(1)$$

as $n \rightarrow \infty$ by Proposition 2.11. Hence $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ is consistent to $(\Pi_n^*; n = 1, 2, \dots)$ although $(\mathcal{M}_{1,n}; n = 1, 2, \dots)$ is not. In this sense, $\mathcal{M}_{2,n}$ is preferable.

As the above example, $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ has better property than $(\mathcal{M}_{1,n}; n = 1, 2, \dots)$. We will call $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ locally consistent to $(\Pi_n; n = 1, 2, \dots)$. We are going to make a formal definition.

Assume $\Theta_n \equiv \Theta \subset \mathbf{R}^p$ and let $d^n \equiv d$ be a usual metric on \mathbf{R}^p . Let $\hat{\theta}_n \in \Theta$ and $\delta_n > 0$ such that $\delta_n \rightarrow 0$. Let

$$\varphi_n : \theta \mapsto \delta_n^{-1}(\theta - \hat{\theta}_n).$$

Let $\mathcal{M}_n = (M_n, e_n)$ be non-random Monte Carlo procedure. For a probability measure Q on Θ , we define a localization Q^* by $Q^*(A) = Q(\varphi_n^{-1}(A)) = Q(\hat{\theta}_n + \delta_n A)$. Let Π_n^* and $e_{n,m}^*(s_\infty)$ be localizations of Π_n and $e_{n,m}(s_\infty)$ with respectively. Then $\mathcal{M}_n^* := (M_n, e_n^*)$ where $e_n^* = (e_{n,m}^*; m = 1, 2, \dots)$ is a non-random Monte Carlo procedure.

Definition 3.8 (Local consistency). $(\mathcal{M}_n; n = 1, 2, \dots)$ is said to be locally consistent to $(\Pi_n; n = 1, 2, \dots)$ if $(\mathcal{M}_n^*; n = 1, 2, \dots)$ is consistent to $(\Pi_n^*; n = 1, 2, \dots)$.

As in the proof of the following proposition, local consistency implies consistency. Moreover, if $(\Pi_n^*; n = 1, 2, \dots)$ is tight, it is consistent to a point mass.

Proposition 3.9. *Assume $(\mathcal{M}_n; n = 1, 2, \dots)$ is locally consistent to $(\Pi_n; n = 1, 2, \dots)$ and $(\Pi_n^*; n = 1, 2, \dots)$ is tight, that is, for any $\epsilon > 0$, there exists a compact set K such that $\limsup_{n \rightarrow \infty} \Pi_n^*(K^c) < \epsilon$. Then $(\mathcal{M}_n; n = 1, 2, \dots)$ is consistent to $(\tilde{\Pi}_n := \delta_{\tilde{\theta}_n}; n = 1, 2, \dots)$*

Proof. By tightness of Π_n^* , $w_\Theta(\Pi_n, \tilde{\Pi}_n) \rightarrow 0$. Write $\Theta_n^* = \varphi_n(\Theta)$. Since

$$\psi \in \text{BL}_1(\Theta) \Leftrightarrow \delta_n^{-1} \psi \in \text{BL}_1(\Theta_n^*),$$

we have $w_\Theta(e_m, \Pi_n) = \delta_n w_{\Theta_n^*}(e_m, \Pi_n^*)$. Hence by local consistency and $\delta_n \rightarrow 0$, we have $R_{m_n}(\mathcal{M}_n, \Pi_n) \rightarrow 0$ for any $m_n \rightarrow \infty$. Therefore by triangular inequality, $R_{m_n}(\mathcal{M}_n, \tilde{\Pi}_n) \leq R_{m_n}(\mathcal{M}_n, \Pi_n) + w(\Pi_n, \tilde{\Pi}_n) \rightarrow 0$. \square

In Example 3.7, $(\mathcal{M}_{1,n}; n = 1, 2, \dots)$ is consistent to $(\Pi_n; n = 1, 2, \dots)$ since

$$R_{m_n}(\mathcal{M}_{1,n}, \Pi_n) = w(Q_{1,n}, \Pi_n) \rightarrow 0.$$

Consistency of $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ comes from local consistency of $(\mathcal{M}_{2,n}; n = 1, 2, \dots)$ by Proposition 3.9.

Remark 3.10. *For the study of non-regular behavior of Monte Carlo procedure, other localization is more natural in some cases. However in the current study, we only use above localization and we do not pursue here for the other possibilities of scaling.*

4 Random Monte Carlo procedure

In this section, we consider random Monte Carlo procedure instead of non-random Monte Carlo procedure. Convergence property of Gibbs sampler will be considered in this framework. Consistency and local consistency are defined as good properties of Monte Carlo procedures.

4.1 Definitions of Monte Carlo and Markov chain Monte Carlo procedure

Let (X, \mathcal{X}, P) be probability space, (S, \mathcal{S}) be a measurable space and (Θ, d) be a complete separable metric space equipped with its Borel σ -algebra Ξ . Let $S^{\mathbb{N}_0}$ be a countable product of S and let $\mathcal{S}^{\otimes \mathbb{N}_0}$ be its σ -algebra. Write an element $S^{\mathbb{N}_0}$ by $s_\infty = (s(0), s(1), \dots)$ and $s_m = (s(0), s(1), \dots, s(m-1)) \in S^m$.

Remark 4.1. *In general, (S, \mathcal{S}) and (Θ, d) may depend on the element of $x \in X$, such that (S_x, \mathcal{S}_x) and (Θ_x, d_x) . Although this dependency is used implicitly, it is not important in this paper. We omit it and simply write (S, \mathcal{S}) and (Θ, d) as above.*

Definition 4.2 (Monte Carlo procedure). *Let M be a probability transition kernel from X to $S^{\mathbf{N}_0}$, that is*

1. $M(x, \cdot)$ is a probability measure on $(S^{\mathbf{N}_0}, \mathcal{S}^{\otimes \mathbf{N}_0})$ for any $x \in X$.
2. $M(\cdot, A_\infty)$ is \mathcal{X} -measurable for any $A_\infty \in \mathcal{S}^{\otimes \mathbf{N}_0}$.

Let e_m be a probability transition kernel from $X \times S^m$ to Θ for $m = 1, 2, \dots$ and $e = (e_m; m = 1, 2, \dots)$. We call $\mathcal{M} = (M, e)$ a Monte Carlo procedure defined on (X, \mathcal{X}, P) on (S, Θ) , or simply, Monte Carlo procedure.

If $S = \Theta$, we call \mathcal{M} , a Monte Carlo procedure defined on (X, \mathcal{X}, P) on Θ . As non-random Monte Carlo procedure, we write $e_m(x, s_\infty)$ or $e_m(x, s_m)$ instead of $e_m(x, s_m, \cdot)$, and we also write $e_m(s_\infty)$ or $e_m(s_m)$ if it does not depend on x . When $S = \Theta$ and $e_m(x, s_m) = m^{-1} \sum_{i=0}^{m-1} \delta_{s(i)}$, we call $e = (e_m; m = 1, 2, \dots)$ a sequence of empirical distribution.

Let μ and K be probability transition kernels from X to S and $X \times S$ to S with respectively. Let M be a probability transition kernel from X to $S^{\mathbf{N}_0}$. When $M(x, \cdot)$ is a Markov measure with initial distribution $\mu(x, \cdot)$, probability transition kernel $K(x, \cdot, \cdot)$, we call M a random Markov measure generated by (μ, K) .

Definition 4.3 (Markov chain Monte Carlo procedure). *When $\mathcal{M} = (M, e)$ is a Monte Carlo procedure and M is a random Markov measure, \mathcal{M} is called a Markov chain Monte Carlo procedure.*

Example 4.4. *Let (X, \mathcal{X}, P) be a probability space and (Y, \mathcal{Y}) be a measurable space and set $S = Y \times \Theta$. Let $P(d\theta|x, y)$ and $P(dy|x, \theta)$ be probability transition kernels from $X \times Y$ to Θ and $X \times \Theta$ to Y . When a random Markov measure M is constructed by a probability transition kernel K defined by*

$$K(x, (y, \theta), d(y^*, \theta^*)) = P(dy^*|x, \theta)P(d\theta^*|x, y^*),$$

then Markov chain Monte Carlo procedure $\mathcal{M} = (M, e)$ is called a Gibbs sampler.

Definition 4.5. *Markov chain Monte Carlo procedure $\mathcal{M} = (M, e)$ and M is called stationary or ergodic if $M(x, \cdot)$ is stationary or ergodic for P -a.e. $x \in X$.*

If M is stationary, for $A \in \Xi$, let $\Pi(x, A) := M(x, \{s_\infty; s(0) \in A\})$. The probability transition kernel Π is called an invariant probability transition kernel for K , M and \mathcal{M} .

4.2 Consistency of Markov chain Monte Carlo procedure

Let $\mathcal{M} = (M, e)$ be a Monte Carlo procedure defined on (X, \mathcal{X}, P) on (S, Θ) and let Π be a probability transition kernel from X to Θ . Let w be a bounded Lipschitz metric on $\mathcal{P}(\Theta)$.

Let

$$W_m(\mathcal{M}(x), \Pi(x)) = \int_{s_\infty \in S^{\mathbf{N}_0}} w(e_m(x, s_\infty), \Pi(x)) M(x, ds_\infty)$$

and

$$R_m(\mathcal{M}, \Pi) = \int_{x \in X} W_m(\mathcal{M}(x), \Pi(x)) P(dx).$$

It is natural to extend the definition of consistency for non-random Monte Carlo procedure to random Monte Carlo procedure as follows.

Definition 4.6 (Consistency). *A Monte Carlo procedure $\mathcal{M} = (M, e)$ defined on (X, \mathcal{X}, P) on (S, Θ) is called consistent to a probability transition kernel Π from X to Θ if $R_m(\mathcal{M}, \Pi) \rightarrow 0$ for $m \rightarrow \infty$.*

Now we consider a sequence of Monte Carlo procedure. Let $(X_n, \mathcal{X}_n, P_n)$ be a probability space and (Θ_n, d^n) be a complete and separable metric space and (S_n, \mathcal{S}_n) be a measurable space for each $n = 1, 2, \dots$. Let w_n be a bounded Lipschitz metric on $\mathcal{P}(\Theta_n)$ defined by a metric d^n . For a Monte Carlo procedure $\mathcal{M}_n = (M_n, e_n)$ for $e_n = (e_{n,m}; m = 1, 2, \dots)$, and probability transition kernel Π_n from X_n to Θ , we define

$$W_m(\mathcal{M}_n(x_n), \Pi_n(x_n)) = \int_{s_\infty \in S_n^{\mathbf{N}_0}} w_n(e_{n,m}(x_n, s_\infty), \Pi_n(x)) M_n(x_n, ds_\infty)$$

and $R_m(\mathcal{M}_n, \Pi_n) = \int W_m(\mathcal{M}_n(x_n), \Pi_n(x_n)) P_n(dx_n)$.

Definition 4.7 (Consistency for random sequence). *A sequence of Monte Carlo procedure $\mathcal{M}_n = (M_n, e_n)$ defined on $(X_n, \mathcal{X}_n, P_n)$ on (S_n, Θ_n) for $n = 1, 2, \dots$ is called consistent to a sequence of probability transition kernels Π_n from X_n to Θ_n for $n = 1, 2, \dots$ if $\lim_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) = 0$ for any $m_n \rightarrow \infty$.*

The definition is different from $\lim_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) = 0$ for certain $m_n \rightarrow \infty$. It should be any $m_n \rightarrow \infty$. For example, a natural Gibbs sampler for simple binomial model (with scaling defined later), for any m_n such that $m_n/n \rightarrow \infty$, the convergence hold. However, it can not take $m_n = \log(n)$. The performance of the Gibbs sampler is very poor in simulation. In this sense, the requirements for “any $m_n \rightarrow \infty$ ” is important. This slow convergence property is called weak consistency, and it will be studied in a separate paper. Fortunately, under regularity condition, Gibbs sampler is consistent under the scaling defined later.

Now we are going to state sufficient conditions for consistency for Markov chain Monte Carlo methods. We can generalize the results of a non-random sequence of Markov chain Monte Carlo procedure to a random sequence. We assume (3.1). Recall that w_∞ is a bounded Lipschitz metric on $\mathcal{P}(\Theta^{\mathbf{N}_0})$.

Proposition 4.8. *Let $\mathcal{M}_n = (M_n, e)$ be a sequence of stationary Monte Carlo procedure defined on $(X_n, \mathcal{X}_n, P_n)$ on Θ with sequence of empirical distribution e for $n = 1, 2, \dots$. Let Π_n be a probability transition kernel from X_n to Θ which is the invariant probability transition kernel of M_n for $n = 1, 2, \dots$. Let $\mathcal{M}_\infty = (M_\infty, e)$ be ergodic, stationary non-random Monte Carlo procedure with invariant probability measure Π_∞ . If*

$$\int w_\infty(M_n(x_n), M_\infty) P_n(dx_n) \rightarrow 0 \tag{4.1}$$

then $(\mathcal{M}_n; n = 1, 2, \dots)$ is consistent to $(\Pi_n; n = 1, 2, \dots)$.

Proof. This is just a direct application of the non-random case. By (3.3),

$$W_{m_n}(\mathcal{M}_n(x_n), \Pi_n(x_n)) \leq W_k(\mathcal{M}_n(x_n), \Pi_n(x_n)) + \frac{k}{m_n}$$

and by triangular inequality,

$$W_k(\mathcal{M}_n(x_n), \Pi_n(x_n)) \leq W_k(\mathcal{M}_n(x_n), \Pi_\infty) + w(\Pi_\infty, \Pi_n(x_n)).$$

By (4.1), $\int w(\Pi_\infty, \Pi_n(x_n)) P_n(dx_n) \rightarrow 0$. Therefore

$$\limsup_{n \rightarrow \infty} R_{m_n}(\mathcal{M}_n, \Pi_n) \leq \limsup_{n \rightarrow \infty} R_k(\mathcal{M}_n, \Pi_\infty).$$

Next we show two convergence properties

$$\lim_{n \rightarrow \infty} R_k(\mathcal{M}_n, \Pi_\infty) \rightarrow R_k(\mathcal{M}_\infty, \Pi_\infty), \quad \lim_{k \rightarrow \infty} R_k(\mathcal{M}_\infty, \Pi_\infty) = 0. \quad (4.2)$$

Define non-random Monte Carlo procedure $\tilde{\mathcal{M}}_n = (\tilde{M}_n, e)$ on Θ by

$$\tilde{M}_n(d\theta_\infty) = \int_{x_n \in X_n} P_n(dx_n) M_n(x_n, d\theta_\infty).$$

Then $R_k(\mathcal{M}_n, \Pi_\infty) = R_k(\tilde{\mathcal{M}}_n, \Pi_\infty)$ and (4.2) becomes convergence of non-random Monte Carlo procedures. Since (4.1) implies $w_\infty(\tilde{M}_n, M_\infty) \rightarrow 0$, the claim follows by Proposition 3.4. \square

Let M be a random Markov measure defined on (X, \mathcal{X}, P) generated by $\mu(x, d\theta)$ and $K(x, \theta, d\theta^*)$. Then we define

$$(\mu \otimes K)(x, d\theta, d\theta^*) = \mu(x, d\theta) K(x, \theta, d\theta^*).$$

The proof of the following proposition is exactly the same as the non-random case. We omit it.

Proposition 4.9. *Stationary Markov chain Monte Carlo procedure $\mathcal{M}_n^i = (M_n^i, e_n^i)$ is defined on $(X_n, \mathcal{X}_n, P_n)$ on Θ with invariant probability transition kernel Π_n^i for $i = 1, 2$ and $n = 1, 2, \dots$. If*

$$\lim_{n \rightarrow \infty} \int \|\Pi_n^1 \otimes K_n^1(x_n, \cdot) - \Pi_n^2 \otimes K_n^2(x_n, \cdot)\| P_n(dx_n) = 0,$$

then $\int w_\infty(M_n^1(x_n), M_n^2(x_n)) P_n(dx_n) \rightarrow 0$.

As the same as the non-random case, we define equivalence for random Monte Carlo procedures.

Definition 4.10 (Equivalence). *Let (X, \mathcal{X}, P) be a probability space and let (Θ, d) be metric space with Borel σ -algebra Ξ , and let (S^i, \mathcal{S}^i) be measurable spaces for $i = 1, 2$. Let $\mathcal{M}^i = (M^i, e^i)$ be Monte Carlo procedure defined on (X, \mathcal{X}, P) on (S^i, Θ) for $i = 1, 2$. Then \mathcal{M}^1 and \mathcal{M}^2 are called equivalent if*

$$R_m(\mathcal{M}^1, \Pi) = R_m(\mathcal{M}^2, \Pi)$$

for any $m \in \mathbf{N}$ and probability transition kernel Π from X to Θ .

4.3 Localization and non-stationarity

In this subsection, we consider two topics, localization and non-stationarity of Monte Carlo procedure. First we define localization for random procedure. For random Monte Carlo procedure, localization is also random. Assume $\Theta_n \equiv \Theta \subset \mathbf{R}^p$ and $d^n \equiv d$ is a usual metric on \mathbf{R}^p . Let $\hat{\theta}_n : X_n \rightarrow \Theta$ be \mathcal{X}_n -measurable map and $\delta_n > 0$ such that $\delta_n \rightarrow 0$. Let

$$\varphi_n : \theta \mapsto \delta_n^{-1}(\theta - \hat{\theta}_n).$$

Let $\mathcal{M}_n = (M_n, e_n)$ be a Monte Carlo procedure. For a probability transition kernel Q from X_n to Θ , let $Q^* = Q_n^*$ be a localization defined by $Q^*(x_n, A) = Q(x_n, \hat{\theta}_n(x_n) + \delta_n A)$ for a Borel set A . Let Π_n^* and $e_{n,m}^*(x_n, s_m, \cdot)$ be localizations defined by $\Pi_n^*(x_n, A) = \Pi_n(x_n, \hat{\theta}_n(x_n) + \delta_n A)$ and $e_{n,m}^*(x_n, s_m, A) = e_{n,m}(x_n, s_m, \hat{\theta}_n(x_n) + \delta_n A)$. Then $\mathcal{M}_n^* := (M_n^*, e_n^*)$ where $e_n^* = (e_{n,m}^*; m = 1, 2, \dots)$ is a Monte Carlo procedure. When $S_n \equiv \Theta$, we may use localization on M_n not on e_n and set $\mathcal{N}_n^* = (M_n^*, e_n)$ by taking $M_n^*(x_n, A_\infty) = M_n(x_n, \hat{\theta}_n + \delta_n A_\infty)$ where $\delta A_\infty = \{(\delta s(0), \delta s(1), \dots); (s(0), s(1), \dots) \in A_\infty\}$. These two localizations \mathcal{M}_n^* and \mathcal{N}_n^* are equivalent.

Definition 4.11. $(\mathcal{M}_n; n = 1, 2, \dots)$ is said to be local consistent to $(\Pi_n; n = 1, 2, \dots)$ if $(\mathcal{M}_n^*; n = 1, 2, \dots)$ is consistent to $(\Pi_n^*; n = 1, 2, \dots)$.

Second, we consider non-stationarity. In other part of the paper, Markov chain Monte Carlo procedure is assumed to be stationary, which is an unrealistic assumption. The choice of the initial probability transition kernel $\mu_n(x_n, d\theta)$ is an important part for designing Monte Carlo method. This choice heavily depends on the structure of model which is more difficult to make a general framework. The following is one possibility which is fundamental proposition for the choice of μ_n .

For $\epsilon > 0$, when two σ -finite measures μ and ν of (E, \mathcal{E}) satisfies $\mu(A) \leq \nu(A) + \epsilon$ for any $A \in \mathcal{E}$, we write $\mu \leq \nu + \epsilon$.

Proposition 4.12. Let $\mathcal{M}_n = (M_n, e_n)$ be stationary Markov chain Monte Carlo where M_n is generated by (Π_n, K_n) . Let $\mathcal{N}_n = (N_n, e_n)$ be another Markov chain Monte Carlo procedure where N_n is generated by (μ_n, K_n) for $\mu_n \neq \Pi_n$. For any $\epsilon > 0$, there exists $c > 0$ such that

$$\limsup_{n \rightarrow \infty} P_n(\{x_n; \mu_n(x_n, \cdot) \leq c\Pi_n(x_n, \cdot) + \epsilon\}^c) < \epsilon. \quad (4.3)$$

Then if $(\mathcal{M}_n; n = 1, 2, \dots)$ is consistent to $(\Pi_n; n = 1, 2, \dots)$, $(\mathcal{N}_n; n = 1, 2, \dots)$ is also consistent to $(\Pi_n; n = 1, 2, \dots)$.

Proof. Take $A_n^\epsilon = \{x_n; \mu_n(x_n, \cdot) \leq c\Pi_n(x_n, \cdot) + \epsilon\}$. If $x_n \in A_n^{\epsilon/2}$,

$$W_m(\mathcal{N}_n(x_n), \Pi_n(x_n)) \leq cW_m(\mathcal{M}_n(x_n), \Pi_n(x_n)) + \epsilon/2.$$

Hence

$$R_m(\mathcal{N}, \Pi_n) \leq P_n((A_n^{\epsilon/2})^c) + cR_m(\mathcal{M}_n, \Pi_n) + \epsilon/2,$$

and take $n \rightarrow \infty$, we have $\limsup_{n \rightarrow \infty} R_{m_n}(\mathcal{N}, \Pi_n) \leq \epsilon$. Hence the claim follows. \square

The meaning of the above proposition becomes clear when we make a localization. It says that with certain regularity of the model and Markov chain Monte Carlo procedure, it is sufficient to find a $n^{1/2}$ -consistent estimator of $\hat{\theta}_n$ to construct a consistent Markov chain Monte Carlo procedure. We illustrate it in the following example.

Example 4.13. Let $\Theta = \mathbf{R}^p$ and $(X_n, \mathcal{X}_n, P_n)$ be a probability space. We prepare some assumptions.

1. $\hat{\theta}_n : X_n \rightarrow \Theta$ is P_n -tight. That is, for any $\epsilon > 0$, there exists a compact set K such that $\limsup_{n \rightarrow \infty} P_n(\{\hat{\theta}_n(x_n) \notin K\}) < \epsilon$.
2. $\mathcal{M}_n = (M_n, e_n)$ is stationary Markov chain Monte Carlo procedure where M_n is generated by (Π_n, K_n) for $n = 1, 2, \dots$. $(\mathcal{M}_n; n = 1, 2, \dots)$ is locally consistent to $(\Pi_n; n = 1, 2, \dots)$ under a map $\varphi_n(x_n) : \theta \mapsto n^{1/2}(\theta - \hat{\theta}_n)$.
3. $I(\theta)$ is a $p \times p$ -positive definite symmetric matrix. It is continuous in θ , that is, for $I(\theta) = (I_{i,j}(\theta)); i, j = 1, \dots, p$,

$$\lim_{n \rightarrow \infty} \sum_{i,j=1}^p |I_{i,j}(\theta_n) - I_{i,j}(\theta)| = 0 \quad (4.4)$$

if $\theta_n \rightarrow \theta$.

4. Transition probability kernel Π_n satisfies

$$\lim_{n \rightarrow \infty} \int \|\Pi_n(x_n, \cdot) - N(\hat{\theta}_n, n^{-1}I(\hat{\theta}_n)^{-1})\| P_n(dx_n) = 0.$$

5. There exists a measurable map $\tilde{\theta}_n : X_n \rightarrow \Theta$ such that $\tau_n := n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)$ is P_n -tight.
6. Q is a probability measure with density q with respect to the Lebesgue measure. The function q is continuous and strictly positive everywhere.

Take

$$\mu_n(x_n, A) = Q(n^{1/2}(A - \tilde{\theta}_n)),$$

Let Π_n^* and μ_n^* be localizations of Π_n and μ_n with respect to $\varphi_n(x_n)$. Then Π_n^* and μ_n^* satisfies (4.3). The proof will be given below. Then if N_n is a random Markov measure defined by (μ_n, K_n) , $\mathcal{N}_n = (N_n, e_n)$ ($n = 1, 2, \dots$) is also locally consistent to $(\Pi_n; n = 1, 2, \dots)$.

Now we prove (4.3) for Π_n^* and μ_n^* . We already know that

$$\lim_{n \rightarrow \infty} \int \|\Pi_n^*(x_n, \cdot) - N(0, I(\hat{\theta}_n)^{-1})\| P_n(dx_n) = 0$$

and for a Borel set A of \mathbf{R}^p ,

$$\mu_n^*(x_n, A) = \mu_n(x_n, \hat{\theta}_n + n^{-1/2}A) = Q(A + \tau_n).$$

Let $B_r := \{\theta \in \mathbf{R}^p; |\theta| < r\}$. For any $\epsilon > 0$, there exists a $R > 0$ such that

$$\limsup_{n \rightarrow \infty} P_n(\hat{\theta}_n \notin B_R) < \epsilon/2, \quad \limsup_{n \rightarrow \infty} P_n(\tau_n \notin B_R) < \epsilon/2, \quad Q(B_R^c) < \epsilon/2.$$

By continuity and positivity of probability distribution functions, there exists constants c_*, c^* such that

$$\inf_{x \in B_{2R}, \theta \in B_R} \phi(x; 0, I(\theta)^{-1}) > c_* > 0, \quad \sup_{x \in B_{3R}} q(x) < c^* < \infty.$$

Take $E_n = \{x_n; \hat{\theta}_n(x_n), \tau(x_n) \in B_R\}$. If $x_n \in E_n$,

$$\mu_n^*(x_n, A) = Q(A + \tau_n) \leq Q((A \cap B_{2R}^c) + \tau_n) + Q((A \cap B_{2R}) + \tau_n),$$

and also

$$\mu_n^*(x_n, A) \leq Q(B_R^c) + c^* \text{Leb}(A \cap B_{2R}) \leq \epsilon/2 + \frac{c^*}{c_*} \int_A \phi(u; 0, I(\hat{\theta})^{-1}) du.$$

Now we set $F_n = \{x_n; \|N(0, I(\hat{\theta}_n)^{-1}) - \Pi_n^*(x_n, \cdot)\| < (c_*/c^*)(\epsilon/2)\}$. Then for $x_n \in E_n \cap F_n$,

$$\mu_n^*(x_n, A) \leq \epsilon + \frac{c^*}{c_*} \Pi_n^*(A).$$

Since $\limsup_{n \rightarrow \infty} P_n((E_n \cap F_n)^c) < \epsilon$, (4.3) holds for Π_n^* and μ_n^* .

Remark 4.14. In this example, μ_n can be computed by the knowledge of Q and $\hat{\theta}_n$. For example, if we can construct a \sqrt{n} -consistent estimator, and if there exists (but can not perform) a locally consistent stationary Markov chain Monte Carlo procedure, then we can construct a non-stationary Markov chain Monte Carlo procedure starting from the \sqrt{n} -consistent estimator with Q .

Remark 4.15. Usually we do not take \sqrt{n} -consistent estimator as a starting point as above remark. On the other hand, we do not take the point to be far from the “center” of target distribution but try to set it to be close to the center. The choice of \sqrt{n} -consistent estimator is not only a recommendation of the choice but also one formalization of the above usual choice.

5 Asymptotic statistics and quadratic mean differentiability

This section provides technical results which will be used later. It may be possible to skip this section and go back if the reader find difficulty reading latter section.

5.1 Quadratic mean differentiability

Let (X, \mathcal{X}) be a measurable space and $\Theta \subset \mathbf{R}^p$ be an open set equipped with Borel σ -algebra Ξ . Let $\mu(dx)$ be σ -finite measure on (X, \mathcal{X}) . Assume also that there exists a $\mathcal{X} \otimes \Xi$ -measurable function $p(x|\theta)$ such that

$$P(dx|\theta) = p(x|\theta)\nu(dx) \quad (5.1)$$

where $P(dx|\theta)$ is a probability transition kernel from Θ to X .

Definition 5.1. $P(dx|\theta)$ is called *quadratic mean differentiable* at θ if there exists a \mathbf{R}^p -valued function $\eta(x|\theta)$ such that for $h \in \mathbf{R}^p$,

$$\int_{x \in X} |\sqrt{p(x|\theta+h)} - \sqrt{p(x|\theta)} - h^T \eta(x|\theta)|^2 \nu(dx) = o(|h|^2)$$

if $|h| \rightarrow 0$.

We call $\eta(x|\theta)$ a quadratic mean derivative of $P(dx|\theta)$ at θ . Let $\tilde{\eta}(x|\theta) = 2\eta(x|\theta)/\sqrt{p(x|\theta)}$, which is called a score function. Let $Z_n(x_n|\theta) = n^{-1/2} \sum_{i=1}^n \tilde{\eta}(x^i|\theta)$ for $x_n = (x^1, \dots, x^n)$. Fisher information matrix $I(\theta)$ is defined by

$$I(\theta) = 4 \int_{x \in X} \eta(x|\theta) \eta(x|\theta)^T \nu(dx) = \int_{x \in X} \tilde{\eta}(x|\theta) \tilde{\eta}(x|\theta)^T P(dx|\theta).$$

Note that $\eta(x|\theta)$ is square integrable with respect to ν if $P(dx|\theta)$ is quadratic mean differentiable at θ and hence $I(\theta)$ exists. Quadratic mean differentiability provides a lot of important results with minimal assumptions. For example if $I(\theta)$ is not singular, the law of $Z_n(x_n|\theta)$ tends to $N(0, I(\theta))$ under $P_n(dx_n|\theta) = \prod_{i=1}^n P(dx^i|\theta)$. See excellent monographs such as [10] and [8]. In this paper, we use the convergence of posterior distribution, which comes from consistency of the posterior distribution and local asymptotic normality of the likelihood.

Let $(X_n, \mathcal{X}_n, P_n(dx_n|\theta))$ be n -th product of $(X, \mathcal{X}, P(dx|\theta))$. Let Λ be a probability measure and $P_n(dx_n) = \int_{\Theta} P_n(dx_n|\theta) \Lambda(d\theta)$. We assume the existence of the probability transition kernel $P_n(d\theta|x_n)$ from X_n to Θ such that

$$P_n(d\theta|x_n) P_n(dx_n) = P_n(dx_n|\theta) \Lambda(d\theta).$$

The following set of assumptions are taken from Theorem 10.1 of [18]. See Theorem 8.1.4 of [9] for other useful set of conditions, in particular, see (A-3,4) of their assumptions.

Assumption 5.2. 1. $P(dx|\theta_1) \neq P(dx|\theta_2)$ if $\theta_1 \neq \theta_2$.

2. $I(\theta)$ is non-singular for any $\theta \in \Theta$ and continuous.

3. For any $\theta_0 \in \Theta$ and $\epsilon > 0$, there exists a sequence $\psi_n : X_n \rightarrow [0, 1]$ such that for $B_\epsilon = \{\theta; |\theta| < \epsilon\}$

$$\lim_{n \rightarrow \infty} \int \psi_n(x_n) P_n(dx_n|\theta_0) = 0, \quad \lim_{n \rightarrow \infty} \sup_{\theta \in B_\epsilon} \int 1 - \psi_n(x_n) P_n(dx_n|\theta) = 0.$$

4. Λ has derivative λ with respect to the Lebesgue measure and λ is continuous and positive.

For fixed $\theta \in \Theta$, write $\tilde{\theta}_n = \theta + n^{-1/2}I(\theta)^{-1}Z_n(x_n|\theta)$. Let $\hat{\theta}_n$ be a central value of $P_n(d\theta|x_n)$.

Proposition 5.3. *Under Assumption 5.2, if $P(dx|\theta)$ is quadratic mean differentiable at any $\theta \in \Theta$, then for any $\theta \in \Theta$, $\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n)$ tends to 0 in $P_n(dx_n|\theta)$ -probability. Moreover,*

$$\lim_{n \rightarrow \infty} \int \|P_n(d\theta|x_n) - N(\hat{\theta}_n, n^{-1}I(\hat{\theta}_n)^{-1})\| P_n(dx_n) = 0. \quad (5.2)$$

Proof. Fix $\theta \in \Theta$. Consider a probability space $(X_\infty, \mathcal{X}_\infty, P_\infty(dx_\infty|\theta))$ which is a countable product of $(X, \mathcal{X}, P(dx|\theta))$. Consider $x_n = (x^1, \dots, x^n)$ as a subsequence of $x_\infty = (x^1, \dots)$. Let $P_n^*(d\theta|x_n)$ be a localization of $P_n(d\theta|x_n)$ by $\theta \mapsto n^{1/2}(\theta - \tilde{\theta}_n)$. Then by Bernstein-von Mises's theorem,

$$\lim_{n \rightarrow \infty} \int \|P_n^*(d\theta|x_n) - N(0, I(\theta)^{-1})\| P_\infty(dx_\infty|\theta) = 0. \quad (5.3)$$

Hence for any subsequence of \mathbf{N} , there exists a further subsequence $n_1 < n_2 < \dots$ such that for $P_\infty(dx_\infty|\theta)$ -a.s. $w(P_{n_i}^*(d\theta|x_{n_i}), N(0, I(\theta)^{-1})) \rightarrow 0$ for $i \rightarrow \infty$. Write $\tau_n = n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)$ the central value of $P_n^*(d\theta|x_n)$. Since the central value is continuous in weak convergence, τ_{n_i} tends to 0 in $P_\infty(dx_\infty|\theta)$ almost surely hence τ_n tends to 0 in $P_\infty(dx_\infty|\theta)$ -probability. Therefore the former claim follows. By continuity of I ,

$$\|N(0, I(\theta)^{-1}) - N(\tau_n, I(\hat{\theta}_n)^{-1})\| \rightarrow 0$$

in $P_\infty(dx_\infty|\theta)$ -probability and by the convergence (5.3), we obtain

$$\int \|P_n^*(d\theta^*|x_n) - N(\tau_n, I(\hat{\theta}_n)^{-1})\| P_n(dx_n|\theta) \rightarrow 0$$

for any θ . Then integrating the right hand side by Λ , the latter claims follows by the dominated convergence theorem. \square

Under $P_n(dx_n|\theta)$ we can construct the following table. This table means that those statistics with the same column is equivalent under $P_n(dx_n|\theta)$, that is, if A_n and B_n are in the same column, $A_n - B_n$ tends in $P_n(dx_n|\theta)$ -probability to 0.

Under $P_n(dx_n|\theta)$, we prefer to use left hand side statistics. Under $P_n(dx_n|\theta)\Lambda(d\theta) = P_n(d\theta|x_n)P_n(dx_n)$, we will use the right hand side. We will use both representation depending on the situation.

Likelihood statistics	Posterior statistics
$\theta, \hat{\theta}_n = \theta + n^{1/2} I(\theta)^{-1} Z_n(x_n \theta)$	$\hat{\theta}_n$: central value of $P_n(d\theta x_n)$
$I(\theta)^{-1} Z_n(x_n \theta)$	$n^{1/2}(\hat{\theta}_n - \theta)$
$I(\theta)$	$I(\hat{\theta}_n)$

Table 1: Equivalent statistics

5.2 Quadratic mean differentiability of marginal model

When we use a Gibbs sampler we usually choose the probability transition kernel (or model) $P(dxdy|\theta)$ from simple parametric family such as exponential family. If $P(dxdy|\theta)$ is an exponential family, quadratic mean differentiability is quite easy to show. On the other hand, quadratic mean differentiability of $P(dx|\theta)$ is sometimes not easy even if $P(dxdy|\theta)$ belongs to an exponential family. In this subsection, under a certain condition, we show that quadratic mean differentiability of $P(dx|\theta)$ comes from that of $P(dxdy|\theta)$.

Let (X, \mathcal{X}) and (Y, \mathcal{Y}) be measurable spaces and $\Theta \subset \mathbf{R}^p$ be an open set with Borel σ -algebra Ξ . Let $\nu(dxdy)$ be σ -finite measure on $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$. Assume that there exists a transition kernel $\nu(dy|x)$ from X to Y such that $\nu(dxdy) = \nu(dx)\nu(dy|x)$ where $\nu(dx) = \int_{y \in Y} \nu(dxdy)$.

Now we forget the assumption for quadratic mean differentiability of $P(dx|\theta)$ and we show the condition from that of $P(dxdy|\theta)$. Assume the existence of a $X \times Y \times \Theta$ -measurable function $p(xy|\theta)$ such that

$$P(dxdy|\theta) = p(xy|\theta)\nu(dxdy).$$

Then (5.1) holds for $p(x|\theta) = \int_{y \in Y} p(xy|\theta)\nu(dy|x)$. Assume $P(dxdy|\theta)$ is quadratic mean differentiable at θ with quadratic mean derivative $\eta(xy|\theta)$. Set

$$\eta(x|\theta) = \begin{cases} \int_{y \in Y} \eta(xy|\theta) \frac{\sqrt{p(xy|\theta)}}{\sqrt{p(x|\theta)}} \nu(dy|x) & \text{if } p(x|\theta) > 0 \\ 0 & \text{otherwise} \end{cases}. \quad (5.4)$$

Proposition 5.4. *Assume $P(dxdy|\theta)$ is quadratic mean differentiable at θ and for any $A \in \mathcal{X} \otimes \mathcal{Y}$, for any $\theta_1, \theta_2 \in \Theta$,*

$$\int_A P(dxdy|\theta_1) > 0 \Leftrightarrow \int_A P(dxdy|\theta_2) > 0.$$

Then $P(dx|\theta)$ is quadratic mean differentiable at θ having quadratic mean derivative $\eta(x|\theta)$ defined in (5.4).

Proof. For $h \in \mathbf{R}^p$, let

$$\begin{aligned} R_h(xy) &= \sqrt{p(xy|\theta+h)} - \sqrt{p(xy|\theta)} - h^T \eta(xy|\theta) \\ r_h(x) &= \sqrt{p(x|\theta+h)} - \sqrt{p(x|\theta)} - h^T \eta(x|\theta). \end{aligned}$$

By assumption, $\int |R_h(xy)|^2 \nu(dxdy) = o(|h|^2)$. We show $\int |r_h(x)|^2 \nu(dx) = o(|h|^2)$. For any $\epsilon > 0$, divide X into three subsets $A_0 = \{x; p(x|\theta) = 0\}$, $A_1 = \{x; p(x|\theta) \in (0, \delta)\}$ and $A_2 = \{x; p(x|\theta) \in [\delta, \infty)\}$ where $\delta = \delta(\epsilon)$ will be defined later.

Fist step We show $\int_{A_0} |r_h(x)|^2 \nu(dx) = 0$. By definition, $p(x|\theta) = 0$ and $\eta(x|\theta) = 0$ for $x \in A_0$. Moreover

$$0 = \int_{x \in A_0} P(dx|\theta) = \int_{A_0 \times Y} P(dxdy|\theta) \Leftrightarrow \int_{A_0 \times Y} P(dxdy|\theta + h) = 0,$$

hence $\int_{A_0} P(dx|\theta + h) = \int_{A_0} p(x|\theta + h) \nu(dx) = 0$. Therefore $r_h(x) = 0$ for ν -a.e. in A_0 which proves the first claim.

Second step We show $\limsup_{n \rightarrow \infty} |h|^{-2} \int_{A_1} |r_h(x)|^2 \nu(dx) \leq \epsilon$ for suitable choice of $\delta > 0$. Set $a^-(x) = \sqrt{p(x|\theta + h)} - \sqrt{p(x|\theta)}$, $A^-(xy) = \sqrt{p(xy|\theta + h)} - \sqrt{p(xy|\theta)}$ and $b(x) = h^T \eta(x|\theta)$ and $B(xy) = h^T \eta(xy|\theta)$. Since $r_h = a^- - b$, by Schwartz's inequality and Minkowskii's inequality, $|r_h(x)|$ is bounded above by

$$|a^-(x)| + |b(x)| \leq \left(\int_Y |A^-(xy)|^2 \nu(dy|x) \right)^{1/2} + \left(\int_Y |B(xy)|^2 \nu(dy|x) \right)^{1/2}.$$

Moreover, since $A^- = R_h + B$ we have

$$|r_h(x)| \leq \left(\int_Y |R_h(xy)|^2 \nu(dy|x) \right)^{1/2} + 2 \left(\int_Y |B(xy)|^2 \nu(dy|x) \right)^{1/2}.$$

Since $|h|^{-1} |B(xy)| \leq |\eta(xy|\theta)|$, for $A_1 = \{p(x|\theta) \in (0, \delta)\}$,

$$\limsup_{|h| \rightarrow 0} |h|^{-2} \int_{A_1} |r_h(x)|^2 \leq 4 \int_{A_1 \times Y} |\eta(xy|\theta)|^2 \nu(dxdy).$$

By dominated convergence theorem, we can take δ small enough to be the right hand side is smaller than ϵ . Hence the second claim follows.

Third step We show $\lim_{n \rightarrow \infty} |h|^{-2} \int_{A_2} |r_h(x)|^2 \nu(dx) = 0$. Let $A^+(xy) = \sqrt{p(xy|\theta + h)} + \sqrt{p(xy|\theta)}$ and $a^+(x) = \sqrt{p(x|\theta + h)} + \sqrt{p(x|\theta)}$. Since $A^- = R_h + B$,

$$\begin{aligned} a^-(x) &= \frac{a^-(x)a^+(x)}{a^+(x)} = \int_Y \frac{A^-(xy)A^+(xy)}{a^+(x)} \nu(dy|x) \\ &= \int_Y \frac{R_h(xy)A^+(xy)}{a^+(x)} \nu(dy|x) + \int_Y \frac{B(xy)A^+(xy)}{a^+(x)} \nu(dy|x) \end{aligned}$$

and denote $s_0(x)$ for the first term of the right hand side. Since $A^+(xy) = A^-(xy) + 2\sqrt{p(xy|\theta)}$, the second term becomes

$$\int_Y \frac{B(xy)A^-(xy)}{a^+(x)} \nu(dy|x) + 2 \int_Y \frac{B(xy)\sqrt{p(xy|\theta)}}{a^+(x)} \nu(dy|x)$$

and denote $s_1(x)$ for the first term of the right hand side. The second term can be simplified by using the relation $\int_Y B(xy) \sqrt{p(xy|\theta)} / \sqrt{p(x|\theta)} \nu(dy|x) = b(x)$. Using this relation, the second term minus $b(x)$ becomes

$$2 \frac{\sqrt{p(x|\theta)}}{a^+(x)} b(x) - b(x) = -b(x) \frac{a^-(x)}{a^+(x)} =: s_2(x).$$

Hence $r_h = a^- - b = s_0 + s_1 + s_2$. The order of the integrals are given in the following table.

	$O(1)$	$O(h ^2)$	$o(h ^2)$
$\nu(dxdy)$ -integral	$ A^+ ^2$	$ A^- ^2, B ^2$	$ R_h ^2$
$\nu(dx)$ -integral		$ a^- ^2, b ^2$	

The table means that for example, since $|A^-|^2$ is categorized in $\nu(dxdy)$ -integral and $O(|h|^2)$,

$$\int_{X \times Y} |A^-(xy)|^2 \nu(dxdy) = O(|h|^2).$$

Since $a^+(x) \geq \delta$ for $x \in A_2$, we do not have to care degeneracy of denominator. By Schwartz's inequality, $(\int_{A_2} |s_0(x)|^2 \nu(dx))^{1/2}$ is bounded above by

$$\delta^{-1} \left(\int |R_h(xy)|^2 \nu(dxdy) \right)^{1/2} \left(\int |A^+(xy)|^2 \nu(dxdy) \right)^{1/2} = o(|h|).$$

Similarly, $(\int_{A_2} |s_1(x)|^2 \nu(dx))^{1/2} = O(|h|^2) = o(|h|)$ and $(\int_{A_2} |s_2(x)|^2 \nu(dx))^{1/2} = O(|h|^2) = o(|h|)$. Since $(\int_{A_2} |r_h(x)|^2 \nu(dx))^{-1/2} \leq \sum_{i=0}^2 (\int_{A_2} |s_i(x)|^2 \nu(dx))^{-1/2} = o(|h|)$, the third claim follows.

Therefore

$$\limsup_{h \rightarrow 0} |h|^{-2} \int |r_h(x)|^2 \nu(dx) = \limsup_{h \rightarrow 0} |h|^{-2} \sum_{i=0}^2 \int_{A_i} |r_h(x)|^2 \nu(dx) \leq \epsilon.$$

This proves the proposition. \square

5.3 Convergence of normalized partial score

Here we assume the condition in Proposition 5.4 and define $\eta(x|\theta)$ as in (5.4). Hence both $P(dxdy|\theta)$ and $P(dx|\theta)$ are quadratic mean differentiable having score functions $\tilde{\eta}(xy|\theta)$ and $\tilde{\eta}(x|\theta)$ and Fisher information matrices $K(\theta)$ and $I(\theta)$ with respectively. Let $\tilde{\eta}(y|x, \theta) = \tilde{\eta}(xy|\theta) - \tilde{\eta}(x|\theta)$ and $J(\theta) = K(\theta) - I(\theta)$. Note that $\int \tilde{\eta}(y|x, \theta) P(dy|x, \theta) = 0$ and

$$\int \tilde{\eta}(y|x, \theta) \tilde{\eta}(y|x, \theta)^T P(dxdy|\theta) = J(\theta).$$

Set

$$Z_n(x_n, y_n | \theta) = n^{-1/2} \sum_{i=1}^n \tilde{\eta}(x^i y^i | \theta), Z_n(x_n | \theta) = n^{-1/2} \sum_{i=1}^n \tilde{\eta}(x^i | \theta)$$

and $Z_n(y_n | x_n, \theta) = Z_n(x_n, y_n | \theta) - Z_n(x_n | \theta)$. We define a probability transition kernel Q from $X_n \times \Theta$ to Θ by

$$Q_n(x_n, \theta, A) = \int_{y_n \in Y_n} 1_A(Z_n(y_n | x_n, \theta)) P_n(dy_n | x_n, \theta). \quad (5.5)$$

Proposition 5.5. *Assume the condition in Proposition 5.4. Suppose that $J(\theta)$ are non-singular and $K(\theta)$ is continuous in θ . Then*

$$\lim_{n \rightarrow \infty} \int w(Q_n(x_n, \theta, \cdot), N(0, J(\theta))) P_n(dx_n | \theta) = 0.$$

Proof. Let $(X_\infty, \mathcal{X}_\infty, P_\infty(dx_\infty | \theta))$ be a countable product of $(X, \mathcal{X}, P(dx | \theta))$. Consider $x_n = (x^1, \dots, x^n)$ as a subsequence of $x_\infty = (x^1, x^2, \dots)$. By the law of large numbers, for $P_\infty(dx_\infty | \theta)$ -a.s.,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \int_{y^i \in Y} \tilde{\eta}(y^i | x^i, \theta) \tilde{\eta}(y^i | x^i, \theta)^T P(dy^i | x^i, \theta) = J(\theta) \quad (5.6)$$

and for $\phi_c(x) = |x|^2 1_{\{|x| > c\}}$, for any $c > 0$, for $P_\infty(dx_\infty | \theta)$ -a.s.,

$$n^{-1} \sum_{i=1}^n \int_{y^i \in Y} \phi_c(\tilde{\eta}(y^i | x^i, \theta)) P(dy^i | x^i, \theta) \rightarrow \int_{y \in Y} \phi_c(\tilde{\eta}(y | x, \theta)) P(dy | x, \theta). \quad (5.7)$$

Let A_∞ be a subset of X_∞ such that the convergences (5.6) and (5.7) holds for $c \in \mathbf{Q}^+ = \{s/t; s, t \in \mathbf{N}\}$, that is, A_∞ is the set satisfying Lindeberg condition. Then $\int_{A_\infty} P_\infty(dx_\infty | \theta) = 1$ and since Lindeberg condition holds,

$$w(Q_n(x_n, \theta, \cdot), N(0, J(\theta))) \rightarrow 0$$

for any $x_\infty \in A_\infty$. Hence the claim follows. \square

6 Local consistency for standard Gibbs sampler

We study local consistency for Gibbs sampler for independent and identically distributed observations. We want to remark that in general, if it does not satisfy regularity conditions, a sequence of standard Gibbs samplers is not always locally consistent. For example, a sequence of usual Gibbs sampler on probit regression model is not locally consistent, which is proved in [7]. This inconsistency partly explains the poor behavior of the Markov chain Monte Carlo procedure.

6.1 Sequence of standard Gibbs sampler

In the latter subsection, we consider local consistency of Gibbs sampler for i.i.d setting. We prepare some notation related to Gibbs sampler in this subsection. First we define standard Gibbs sampler for a general space. After that, we define standard Gibbs sampler for more specific situation, more precisely, for i.i.d. setting.

Let (Θ, d) be a complete separable metric space equipped with a Borel σ -algebra Ξ and let (X, \mathcal{X}, P) and (Y, \mathcal{Y}) be a probability space and a measurable space with respectively. Set $(S, \mathcal{S}) := (Y, \mathcal{Y}) \otimes (\Theta, \Xi)$. Write an element of S by $s = (y, \theta)$. Assume that there are probability transition kernels $P(dy|x, \theta)$, $P(d\theta|x, y)$, $P(d\theta|x)$ and $P(dy|x)$ such that for P -a.s. x ,

$$P(dy|x, \theta)P(d\theta|x) = P(d\theta|x, y)P(dy|x). \quad (6.1)$$

When the above relation holds, we define a probability transition kernel \overline{K} from $X \times S$ to S and another probability transition kernel K from $X \times \Theta$ to Θ such that

$$\overline{K}(x, s, ds^*) = P(dy^*|x, \theta)P(d\theta^*|x, y^*) \quad (6.2)$$

for $s = (y, \theta)$ and $s^* = (y^*, \theta^*)$ and

$$\overline{K}(x, \theta, d\theta^*) = \int_{y \in Y} P(dy|x, \theta)P(d\theta^*|x, y). \quad (6.3)$$

Note that $\overline{K}(x, s, ds^*)$ does not depend on y .

We show that $P(d\theta|x)$ is the invariant probability transition kernel of K . See the end of Section 4.1 for the definition of the invariant probability transition kernel.

Proposition 6.1. *Under (6.1), the invariant probability transition kernel of K is $\Pi(x, d\theta) = P(d\theta|x)$.*

Proof. Without loss of generality, we can assume (6.1) for any $x \in X$. By definition, for any $A \in \Xi$,

$$\int_{\theta \in \Theta, \theta^* \in A} \Pi(x, d\theta)K(\theta, d\theta^*) = \int_{\theta \in \Theta, y \in Y, \theta^* \in A} P(d\theta|x)P(dy|x, \theta)P(d\theta^*|x, y).$$

Using (6.1), we can integrate out θ and then using (6.1) again, we can also integrate out y . This calculation yields

$$(\Pi \circ K)(x, A) := \int_{\theta \in \Theta, \theta^* \in A} \Pi(x, d\theta)K(\theta, d\theta^*) = \int_{\theta^* \in A} P(d\theta^*|x) = \Pi(x, A).$$

Since Ξ is countably generated, there exists a subset Σ having countable number of elements which generate Ξ . Then

$$\tilde{X} := \{x \in X; (\Pi \circ K)(x, A) = \Pi(x, A) \ (\forall A \in \Sigma)\}$$

has measure 1 under $P(dx)$. Now fix an element x of \tilde{X} . Then a subset of Ξ defined by

$$\tilde{\Xi} = \tilde{\Xi}_x := \{A \in \Xi; (\Pi \circ K)(x, A) = \Pi(x, A)\}$$

is σ -algebra which contains Σ . Hence $\Xi = \tilde{\Xi}$ for any $x \in \tilde{X}$. This means that if $x \in \tilde{X}$, then $\Pi \circ K(x, \cdot) = \Pi(x, \cdot)$ and $\int_{\tilde{X}} P(dx) = 1$, that is, Π is an invariant probability transition kernel of K . \square

Now we define standard Gibbs sampler. Let $\bar{e} = (\bar{e}_m; m = 1, 2, \dots)$ be a sequence of probability transition kernels \bar{e}_m defined by

$$\bar{e}_m(s_m, \cdot) = \frac{1}{m} \sum_{i=0}^{m-1} \delta_{\theta(i)}$$

where $s_m = (s(0), \dots, s(m-1)) \in S^m$ and $s(i) = (y(i), \theta(i))$. We call \bar{e} a sequence of empirical distribution on Θ .

Definition 6.2. Assume that there are probability transition kernels

$$P(dy|x, \theta), P(d\theta|x, y), P(d\theta|x), P(dy|x)$$

such that (6.1) hold. Let M be a random Markov measure generated by $(\bar{\mu}, \bar{K})$ for $\bar{\mu}(x, ds) = P(d\theta|x)P(dy|x, \theta)$ and \bar{K} defined by (6.2). Then $\bar{M} = (\bar{M}, \bar{e})$ is called a standard Gibbs sampler on (X, \mathcal{X}, P) on (S, Θ) when \bar{e} is a sequence of empirical distribution on Θ .

Now we concentrate on Gibbs sampler for more specific setting, i.i.d. setting. Let (X, \mathcal{X}) , (Y, \mathcal{Y}) be measurable space and (Θ, d) be a complete and separable metric space with a Borel σ -algebra Ξ . A probability measure Λ is defined on (Θ, Ξ) . Let $P(dxdy|\theta)$ be a probability transition kernel from Θ to $X \times Y$ such that there exists probability transition kernels $P(dy|x, \theta)$ and $P(dx|\theta)$ satisfying $P(dxdy|\theta) = P(dy|x, \theta)P(dx|\theta)$. A sequence of standard Gibbs sampler will be constructed by these probability measure and probability transition kernel.

Let (X_n, \mathcal{X}_n) and (Y_n, \mathcal{Y}_n) be n -th products of (X, \mathcal{X}) and (Y, \mathcal{Y}) . Write their elements by $x_n = (x^1, \dots, x^n)$ and $y_n = (y^1, \dots, y^n)$ with respectively. We define probability transition kernels

$$P_n(dx_n|\theta) = \prod_{i=1}^n P(dx^i|\theta), \quad P_n(dx_n dy_n|\theta) = \prod_{i=1}^n P(dx^i dy^i|\theta),$$

$$P_n(dy_n|x_n, \theta) = \prod_{i=1}^n P(dy^i|x^i, \theta),$$

and probability measures

$$P_n(dx_n) = \int_{\Theta} P_n(dx_n|\theta) \Lambda(d\theta), \quad P_n(dx_n dy_n) = \int_{\Theta} P_n(dx_n dy_n|\theta) \Lambda(d\theta).$$

Assume that there are probability transition kernels $P_n(d\theta|x_n, y_n)$ and $P_n(d\theta|x_n)$ satisfying

$$P_n(d\theta|x_n, y_n)P_n(dx_n dy_n) = P_n(dx_n dy_n|\theta)\Lambda(d\theta) \quad (6.4)$$

and

$$P_n(d\theta|x_n)P_n(dx_n) = P_n(dx_n|\theta)\Lambda(d\theta). \quad (6.5)$$

Moreover, we assume (6.1) for these transition kernels, that is,

$$P_n(dy_n|x_n, \theta)P_n(d\theta_n|x_n) = P_n(d\theta|x_n, y_n)P_n(dy_n|x_n).$$

Note that this relation is automatically satisfied if \mathcal{X} is countably generated. Let $(S_n, \mathcal{S}_n) = (Y_n, \mathcal{Y}_n) \otimes (\Theta, \Xi)$.

Definition 6.3. Assume (6.1, 6.4, 6.5). Let $\overline{\mathcal{M}}_n = (\overline{M}_n, \overline{e}_n)$ be a standard Gibbs sampler on $(X_n, \mathcal{X}_n, P_n(dx_n))$ on (S_n, Θ) defined by

$$P_n(dy_n|x_n, \theta), P_n(d\theta|x_n, y_n), P_n(d\theta|x_n), P_n(dy_n|x_n).$$

Then $(\overline{\mathcal{M}}_n; n = 1, 2, \dots)$ is called a sequence of standard Gibbs sampler generated by $P(dx dy|\theta)$ and Λ .

Later, we will consider analysis of $\overline{\mathcal{M}}_n$. For that purpose, it is convenient to consider alternative equivalent Monte Carlo procedure. Let e be a sequence of empirical distribution and let a probability transition kernel K_n from $X_n \times \Theta_n$ to Θ_n be

$$K_n(x_n, \theta, d\theta^*) = \int_{y_n \in Y_n} P_n(dy_n|x_n, \theta)P_n(d\theta|x_n, y_n). \quad (6.6)$$

Let M_n be a random Markov measure generated by (Π_n, K_n) where $\Pi_n(x_n, d\theta) = P_n(d\theta|x_n)$. Then $\mathcal{M}_n = (M_n, e)$ is equivalent to $\overline{\mathcal{M}}_n$. We refer to the Markov chain Monte Carlo procedure \mathcal{M}_n by minimal equivalent Markov chain Monte Carlo procedure.

6.2 Approximation of the standard Gibbs sampler

In this subsection, we fix $\theta_0 \in \Theta \subset \mathbf{R}^p$ and all arguments are under $P_n(dx_n|\theta_0)$ and $P_n(dx_n dy_n|\theta_0)$. We assume the same condition as Section 5.3. Table 2 for equivalent statistics is useful, which is an extension of Table 1.

Write the central values of $P_n(d\theta|x_n y_n)$ and $P_n(d\theta|x_n)$ by $\hat{\theta}_n(x_n, y_n)$ and $\hat{\theta}_n(x_n)$ with respectively. In the following, we write $A \equiv^a B$ if $A - B$ tends in $P_n(dx_n dy_n|\theta_0)$ -probability to 0. Then by Table 2 and by $I_\theta^{-1} = K_\theta^{-1}(J_\theta + I_\theta)I_\theta^{-1} = K_\theta^{-1}J_\theta I_\theta^{-1} + K_\theta^{-1}$,

$$\begin{aligned} n^{1/2}(\hat{\theta}_n(x_n, y_n) - \hat{\theta}_n(x_n)) &\equiv^a K_{\theta_0}^{-1}Z_n(x_n, y_n|\theta_0) - I_{\theta_0}^{-1}Z_n(x_n|\theta_0) \\ &\equiv^a K_{\theta_0}^{-1}Z_n(y_n|x_n, \theta_0) - K_{\theta_0}^{-1}J_{\theta_0}I_{\theta_0}^{-1}Z_n(x_n|\theta_0) \\ &\equiv^a K_{\theta_0}^{-1}Z_n(y_n|x_n, \theta_0) + n^{1/2}K_{\theta_0}^{-1}J_{\theta_0}(\theta_0 - \hat{\theta}_n(x_n)) \end{aligned}$$

Likelihood statistics	Posterior statistics
$\theta_0, \hat{\theta}_n(x_n), \hat{\theta}_n(x_n, y_n)$	$\hat{\theta}_n(x_n), \hat{\theta}_n(x_n, y_n)$
$I(\theta_0)^{-1} Z_n(x_n \theta_0)$	$n^{1/2}(\hat{\theta}_n(x_n) - \theta_0)$
$K(\theta_0)^{-1} Z_n(x_n, y_n \theta_0)$	$n^{1/2}(\hat{\theta}_n(x_n, y_n) - \theta_0)$
$I(\theta_0)$	$I(\hat{\theta}_n)$
$K(\theta_0)$	$K(\hat{\theta}_n)$
$J(\theta_0)$	$J(\hat{\theta}_n)$

Table 2: Equivalent statistics

where $L_{\theta_0} = L(\theta_0)$ for $L = I, J$ and K . By Proposition 5.5, the law of $Z_n(y_n | x_n, \theta_0)$ tends to $N(0, J(\theta_0))$. Hence, formally, we replace $Z_n(y_n | x_n, \theta_0)$ by ξ_1 which follows $N_p(0, I)$ (where I is the $p \times p$ identity matrix), that is,

$$n^{1/2}(\hat{\theta}_n(x_n, y_n) - \hat{\theta}_n(x_n)) \sim K_{\theta_0}^{-1} J_{\theta_0}^{1/2} \xi_1 + n^{1/2} K_{\theta_0}^{-1} J_{\theta_0}(\theta_0 - \hat{\theta}_n(x_n))$$

where \sim means “similar” in certain sense (just a formal argument). Since $P_n(d\theta | x_n, y_n)$ tends to $N(\hat{\theta}_n(x_n, y_n), n^{-1} K(\theta_0)^{-1})$, the realization θ^* from $P_n(d\theta | x_n, y_n)$ satisfies

$$n^{1/2}(\theta^* - \hat{\theta}_n(x_n, y_n)) \sim K_{\theta_0}^{-1/2} \xi_2$$

where ξ_2 follows $N_p(0, I)$. Hence

$$n^{1/2}(\theta^* - \hat{\theta}_n(x_n)) \sim K_{\theta_0}^{-1/2} \xi_2 + K_{\theta_0}^{-1} J_{\theta_0}^{1/2} \xi_1 + n^{1/2} K_{\theta_0}^{-1} J_{\theta_0}(\theta_0 - \hat{\theta}_n(x_n))$$

where ξ_1 and ξ_2 follows $N_p(0, I)$ independently. Therefore we approximate $K_n(x_n, \theta_0, d\theta^*)$ defined by (6.6) by

$$N(\hat{\theta}_n(x_n) + K_{\theta_0}^{-1} J_{\theta_0}(\theta_0 - \hat{\theta}_n(x_n)), n^{-1} K_{\theta_0}^{-1} + n^{-1} K_{\theta_0}^{-1} J_{\theta_0} K_{\theta_0}^{-1}).$$

By replacing I, J, K at θ_0 by $\hat{I}, \hat{J}, \hat{K} := I, J, K$ at $\hat{\theta}_n(x_n)$

$$\tilde{K}_n(x_n, \theta_0, \cdot) := N(\hat{\theta}_n(x_n) + \hat{K}^{-1} \hat{J}(\theta_0 - \hat{\theta}_n(x_n)), n^{-1} \hat{K}^{-1} + n^{-1} \hat{K}^{-1} \hat{J} \hat{K}^{-1}).$$

Since $\Pi_n(x_n, d\theta) = P_n(d\theta | x_n)$ is approximated by $N(\hat{\theta}_n(x_n), n^{-1} \hat{I})$,

$$\tilde{\Pi}_n(x_n, \cdot) := N(\hat{\theta}_n(x_n), n^{-1} \hat{I}).$$

We approximate M_n by a random Markov measure \tilde{M}_n generated by $(\tilde{\Pi}_n, \tilde{K}_n)$.

6.3 Local consistency of the standard Gibbs sampler

We study local consistency of a standard Gibbs sampler. Before stating the result, we make one remark for initial probability transition kernel. For fixed observation, the standard Gibbs sampler uses the posterior distribution as an

initial distribution, which is unrealistic. However as mentioned in Section 4.3, we can replace the initial distribution by small perturbation from $n^{1/2}$ -consistent estimator as in Example 4.13.

Let $\hat{\theta}_n(x_n)$ be a central value of $P_n(d\theta|x_n)$. We consider localization by $\theta \mapsto n^{1/2}(\theta - \hat{\theta}_n(x_n))$.

Theorem 6.4. *Assume the condition in Proposition 5.4 and Assumption 5.2. Suppose that $I(\theta)$ and $J(\theta)$ are non-singular and $I(\theta), K(\theta)$ are continuous in θ . Then the standard Gibbs sampler $(\bar{\mathcal{M}}_n; n = 1, 2, \dots)$ is locally consistent to $(\Pi_n; n = 1, 2, \dots)$*

Proof. It is sufficient to study a sequence of minimal equivalent Markov chain Monte Carlo procedure $\mathcal{M}_n = (M_n, e)$ defined after Definition 6.3. First we show

$$\lim_{n \rightarrow \infty} \int \|\Pi_n \otimes K_n(x_n, \cdot) - \tilde{\Pi}_n \otimes \tilde{K}_n(x_n, \cdot)\| P_n(dx_n) = 0 \quad (6.7)$$

where $\tilde{\Pi}_n$ and \tilde{K}_n are defined in the previous subsection. By triangular inequality, we have

$$\|\Pi_n \otimes K_n - \tilde{\Pi}_n \otimes \tilde{K}_n\| \leq \|\Pi_n \otimes K_n - \Pi_n \otimes \tilde{K}_n\| + \|\Pi_n \otimes \tilde{K}_n - \tilde{\Pi}_n \otimes \tilde{K}_n\|$$

and the second term of the right hand side is bounded by $\|\Pi_n - \tilde{\Pi}_n\|$ which tends in $P_n(dx_n)$ -probability to 0 by Proposition 5.3. Since $\Pi_n(x_n, d\theta) = P_n(d\theta|x_n)$, the first term integrated by $P_n(dx_n)$ is bounded by

$$\int \|(K_n - \tilde{K}_n)(x_n, \theta, \cdot)\| P_n(d\theta|x_n) P_n(dx_n) = \int \|(K_n - \tilde{K}_n)(x_n, \theta, \cdot)\| P_n(dx_n|\theta) \Lambda(d\theta).$$

We fix $\theta \in \Theta$ and consider the convergence of the integrand of Λ . Use the likelihood statistics in the sense of Table 2. To show this convergence, we make two probability transition kernels L_n and \tilde{L}_n and consider inequality

$$\|K_n - \tilde{K}_n\| \leq \|K_n - L_n\| + \|L_n - \tilde{L}_n\| + \|\tilde{L}_n - \tilde{K}_n\|. \quad (6.8)$$

First, we construct \tilde{L}_n by

$$\tilde{L}_n(x_n, \theta, \cdot) := N(\tilde{\theta}_n(x_n) - n^{-1/2} K_\theta^{-1} J_\theta I_\theta^{-1} Z_n(x_n|\theta), n^{-1} K_\theta^{-1} + n^{-1} K_\theta^{-1} J_\theta K_\theta^{-1})$$

where $\tilde{\theta}_n(x_n, y_n) = \theta + n^{-1/2} K_\theta^{-1} Z_n(x_n, y_n|\theta)$. Then the integral of the last term of the right hand side of (6.8) tends to 0 since the differences between \tilde{L}_n and \tilde{K}_n are made by asymptotically equivalent statistics (see Table 2). Second, we make L_n by

$$L_n(x_n, \theta, d\theta^*) = \int_{y_n \in Y_n} P_n(dy_n|x_n, \theta) \phi(\theta^*; \tilde{\theta}_n(x_n, y_n), n^{-1} K_\theta^{-1}) d\theta^*.$$

Then the first term of the right hand side of (6.8) integrated by $P_n(dx_n|\theta)$ is bounded by

$$\int \|P_n(d\theta|x_n, y_n) - N(\tilde{\theta}_n(x_n, y_n), n^{-1} K_\theta^{-1})\| P_n(dx_n dy_n|\theta) \rightarrow 0$$

which is a consequence of Bernstein-von Mises's theorem for $P(dxdy|\theta)$. Third, we consider the middle term of the right hand side of (6.8). Now we make a localization by

$$\theta^* \mapsto n^{1/2}(\theta^* - \tilde{\theta}(x_n) + n^{-1/2} K_\theta^{-1} J_\theta I_\theta^{-1} Z_n(x_n|\theta)) = n^{1/2}(\theta^* - \theta - n^{-1/2} K_\theta^{-1} Z_n(x_n|\theta)).$$

Then localizations of \tilde{L}_n and L_n are $\tilde{L}_n^*(x_n, \theta, \cdot) = N(0, K_\theta^{-1} + K_\theta^{-1} J_\theta K_\theta^{-1})$ and

$$L_n^*(x_n, \theta, du) = \int_{y_n \in Y_n} P_n(dy_n|x_n, \theta) \phi(u; K_\theta^{-1} Z_n(y_n|x_n, \theta), K_\theta^{-1}) du.$$

Note that

$$\tilde{L}_n^*(x_n, \theta, du) = \int_{v \in \mathbf{R}^p} \phi(v; 0, J_\theta) dv \phi(u; K_\theta^{-1} v, K_\theta^{-1}) du.$$

Take $B_r = \{x \in \mathbf{R}^p; |x| < r\}$ and set

$$\psi_u(v) = \phi(u; K_\theta^{-1} v, K_\theta^{-1})$$

which is a Lipschitz continuous function with Lipschitz constant $c(\theta) > 0$. Then for $M > 0$, $\|(L_n^* - \tilde{L}_n^*)(x_n, \theta, \cdot)\|$ is bounded above by

$$\tilde{L}_n^*(x_n, \theta, B_M^c) + \int_{u \in B_M} \left| \int_{v \in \mathbf{R}^p} \psi_u(v) Q_n(x_n, \theta, dv) - \int_{v \in \mathbf{R}^p} \psi_u(v) \phi(v; 0, J_\theta) dv \right| du$$

where Q_n is defined by (5.5). The latter is bounded above by

$$\text{Leb}(B_M) c(\theta) w(Q_n(x_n, \theta, dv), N(0, J_\theta)). \quad (6.9)$$

Since $\tilde{L}_n^*(x_n, \theta, du)$ does not depend on x_n , we can take M large enough to be $\tilde{L}_n^*(x_n, \theta, B_M^c) \leq \epsilon$ (say) and integral of (6.9) tends to 0 by Proposition 5.5. Hence $\int \|L_n - \tilde{L}_n\| P_n(dx_n|\theta) \rightarrow 0$ for any θ , which completes the proof of (6.7).

Now we make a localization $\theta \mapsto n^{1/2}(\theta - \hat{\theta}(x_n))$. Unfortunately, since $\tilde{\Pi}_n^*$ and \tilde{K}_n^* are random, we can not directly use Propositions 4.8 and 4.9 to conclude local consistency of \mathcal{M}_n . However since $P_n(dx_n) = P_n(dx_n|\theta) \Lambda(d\theta)$,

$$\lim_{n \rightarrow \infty} \int \|\Pi_n^* \otimes K_n^*(x_n, \cdot) - \tilde{\Pi}_n^* \otimes \tilde{K}_n^*(x_n, \cdot)\| P_n(dx_n|\theta) \rightarrow 0 \quad (6.10)$$

in Λ -probability, and for each θ , we can replace $\tilde{\Pi}_n^*(x_n, \cdot)$ and $\tilde{K}_n^*(x_n, \cdot)$ by non-random kernels $\hat{\Pi}$ and \hat{K} where

$$\hat{\Pi}(\cdot) = N(0, I^{-1}(\theta)), \quad \hat{K}(u, \cdot) = N(K^{-1}(\theta) J(\theta) u, K^{-1}(\theta) + K^{-1}(\theta) J(\theta) K^{-1}(\theta)).$$

Therefore,

$$\lim_{n \rightarrow \infty} \int \|\Pi_n^* \otimes K_n^*(x_n, \cdot) - \hat{\Pi} \otimes \hat{K}\| P_n(dx_n|\theta) \rightarrow 0 \quad (6.11)$$

in Λ -probability. Fix $m_n \rightarrow \infty$. For any subsequence of \mathbf{N} , there is a further subsequence $n_1 < n_2 < \dots$ such that the above convergence holds for Λ -a.e. θ replacing n by n_i . Then we can apply Propositions 4.8 and 4.9 to conclude consistency of $\mathcal{M}_{n_i}^* = (M_{n_i}^*, e_{n_i})$ under $P_n(d\theta|x_n)$, that is,

$$\int_{x_{n_i} \in X_{n_i}} \int_{\theta_\infty} w(e_{m_{n_i}}(\theta_\infty), \Pi_{n_i}^*(x_{n_i})) M_{n_i}^*(x_{n_i}, d\theta_\infty) P_{n_i}(dx_{n_i}|\theta) \rightarrow 0$$

for Λ -a.e. θ where M_n^* is a localization of M_n . Since the convergence is true for some subsequence choosing from any subsequence of \mathbf{N} , we have

$$\int_{x_n \in X_n} \int_{\theta_\infty} w(e_{m_n}(\theta_\infty), \Pi_n^*(x_n)) M_n^*(x_n, d\theta_\infty) P_n(dx_n) \rightarrow 0$$

which means consistency of \mathcal{M}_n^* , that is the desired conclusion. \square

7 Discussion

In this paper, we defined Monte Carlo procedure and Markov chain Monte Carlo procedure as a set of probability measure on $S^{\mathbf{N}_0}$ with a sequence of map from finite product of S to Θ . In particular, we studied local consistency for a sequence of standard Gibbs sampler under regularity conditions. This property is a good behavior of a sequence of Markov chain Monte Carlo procedure.

What we did not discuss in this paper was the following.

1. Research for poor behavior analysis of a sequence of Markov chain Monte Carlo procedure. In fact, if the sequence has a good property, we do not have to tune up the algorithm since we already have a good Monte Carlo procedure. The poor behavior can be studied degeneracy and local degeneracy of Markov chain Monte Carlo procedure. Moreover, we can define a rate of convergence. This research will be studied in a separate work such as for mixture model and categorical data model.
2. Research for constructing new Monte Carlo procedure. Unfortunately, the analysis in the paper is for usual Monte Carlo procedures. We believe that these analysis is useful for constructing new Monte Carlo procedures. The paper [6] shows one possibility.
3. We do not consider point estimation but posterior approximation. This is just for simplicity. Let $(\mathcal{M}_n = (M_n, e); n = 1, 2, \dots)$ be consistent to $(\Pi_n; n = 1, 2, \dots)$ where e is a sequence of empirical distribution. Then it is easy to show that if M_n is stationary, then $m^{-1} \sum_{i=0}^{m-1} \theta(i)$ tends to $\int \theta \Pi_n(x_n, d\theta)$ when $(\Pi_n^1(x_n, d\theta) := |\theta| \Pi_n(x_n, d\theta); n = 1, 2, \dots)$ is tight. When M_n is not stationary, and (4.3) holds, the same conclusion holds if we make burn-in. Note that without burn-in, it may not be true.

Acknowledgements

This is, essentially, the second part of the author's Ph.D. thesis. The author wishes to express his thanks to the Ph.D. supervisor, Prof. Nakahiro Yoshida for his several helpful comments and suggestions.

References

- [1] Yves F. Atchade and Francois Perron. Improving on the independent Metropolis-Hastings algorithm. *Statistica Sinica*, 15:3–18, 2005.
- [2] Alexandre Belloni and Victor Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *Annals of Statistics*, 37(4):2011–2055, 2009.
- [3] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs Sampling, Exponential Families and Orthogonal Polynomials (with discussion). *Statistical Science*, 23(2):151–200, 2008.
- [4] J. L. Doob. *Stochastic Processes*. John Wiley and Sons, Inc., 1953.
- [5] Kiyosi Ito. *Stochastic Processes: Lectures Given at Aarhus University*. Springer, 2004.
- [6] Kengo Kamatani. Metropolis-Hastings Algorithm for Mixture Model and its Weak Convergence. In Gilbert Lechevallier, Yves; Saporta, editor, *Proceedings of COMPSTAT'2010*, volume eBook, pages 1175–1182, 2010.
- [7] Kengo Kamatani. Local degeneracy of Markov chain Monte Carlo methods. *Arxiv*, 2011.
- [8] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1 edition, 1986.
- [9] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics : some basic concepts*. New York ; Tokyo : Springer-Verlag, 2nd edition, 2000.
- [10] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.
- [11] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.
- [12] Soren Feodor Nielsen. The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.
- [13] Esa Nummelin. *General irreducible Markov chains and nonnegative operators*. Number 83 in Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1984.

- [14] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.
- [15] Albert N. Shiryaev. *Probability*. Springer, 2nd edition, 1995.
- [16] Ingrid Svensson and Sara Sjöstedt de Luna. Asymptotic properties of a stochastic em algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference*, 140:117–127, 2010.
- [17] L. Tierney. Markov Chains for Exploring Posterior Distributions (with discussion). *The Annals of Statistics*, 22(4):1701–1762, 1994.
- [18] A.W. van der Vaart. *Asymptotic statistics*. Cambridge ; New York : Cambridge University Press, 1998.